On the Role of Noise in the Sample Complexity of Learning Recurrent Neural Networks: Exponential Gaps for Long Sequences

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider the class of noisy multi-layered sigmoid recurrent neural networks with w (unbounded) weights for classification of sequences of length T, where independent noise distributed according to $\mathcal{N}(0, \sigma^2)$ is added to the output of each neuron in the network. Our main result shows that the sample complexity of PAC learning this class can be bounded by $O(w \log(T/\sigma))$. For the non-noisy version of the same class (i.e., $\sigma = 0$), we prove a lower bound of $\Omega(wT)$ for the sample complexity. Our results indicate an exponential gap in the dependence of sample complexity on T for noisy versus non-noisy networks. Moreover, given the mild logarithmic dependence of the upper bound on $1/\sigma$, this gap still holds even for numerically negligible values of σ.

1 Introduction

Recurrent Neural Networks (RNNs) are effective tools for processing sequential data. They are used in numerous applications such as speech recognition (Graves et al., 2013), computer vision (Karpathy and Fei-Fei, 2015), translation (Sutskever et al., 2014), modeling dynamical systems (Hardt et al., 2018) and time series (Qin et al., 2017). Recurrent models allow us to design classes of predictors that can be applied to (i.e., take input values from) sequences of arbitrary length. For processing a sequence of T elements, a predictor f (e.g., a neural network) "consumes" the input elements one by one, generating an output at each step. This output is then used in the next step (as another input to f along with the next element in the input sequence). Defining recurrent models formally takes some effort, and we relegate it to the next sections. In short, the function f is (recursively) applied T times in order to generate the ultimate outcome.

Let us fix a base class $\mathcal{F}_w$ of all multi-layered feed-forward sigmoid neural networks with w weights. We can create a recurrent version of this class, which we will denote by REC[$\mathcal{F}_w, T$], for classifying sequences of length T. One can study the sample complexity of PAC learning REC[$\mathcal{F}_w, T$] with respect to different loss functions. Koiran and Sontag (1998) studied the binary-valued version of this class by applying a threshold function at the end, and proved a lower bound of $\Omega(wT)$ for its VC dimension.

There has also been efforts for proving upper bounds on the sample complexity of PAC learning REC[$\mathcal{F}, T$] for various base classes $\mathcal{F}$ and different loss functions. Given the above lower bound, a gold standard has been achieving a linear dependence on T in the upper bound. Koiran and Sontag (1998) proved an upper bound of $O(w^4T^2)$ on the VC dimension of REC[$\mathcal{F}_w, T$] discussed above. More recent papers have considered the more realistic setting of classification with continuous-valued RNNs, e.g., by removing the threshold function and using a bounded Lipschitz surrogate loss. In this

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

setting, Zhang et al. (2018) proved an upper bound of $\widetilde{O}(T^4 w \|W\|^{O(T)})$ on the sample complexity[1] where $\|W\|$ is the spectral norm of the network. Chen et al. (2020) improved over this result by proving an upper bound of $\widetilde{O}(Tw\|W\|^2 \min\{\sqrt{w}, \|W\|^{O(T)}\})$. These bounds get close to the gold standard when the spectral norm of the network satisfies $\|W\| \leq 1$.

The above upper bounds are proved by simply "unfolding" the recurrence, effectively substituting the recurrent class REC[$\mathcal{F}_w, T$] with the (larger) class of T-fold compositions $\mathcal{F}_w \circ \mathcal{F}_w \ldots \circ \mathcal{F}_w$. These unfolding techniques do not exploit the fact that the function f (that is applied recursively for T steps to compute the output of the network) is fixed across all the T steps. Consequently, the resulting sample complexity has (super-)linear dependence on T. Therefore, we would need a prohibitively large sample size for training recurrent models for classifying very long sequences. Nevertheless, this dependence is inevitable in light of the of lower bound of Koiran and Sontag (1998). Or is it?

In this paper, we consider a related class of *noisy* recurrent neural networks, REC[$\widetilde{\mathcal{F}_w^\sigma}, T$]. The hypotheses in this class are similar to those in REC[$\mathcal{F}_w, T$], except that outputs of (sigmoid) activation functions are added with independent Gaussian random variables, $\mathcal{N}(0, \sigma^2)$. Our main result demonstrates that, remarkably, the noisy class can be learned with a number of samples that is only logarithmic with respect to T.

Theorem 1 (Informal version of Theorem 15)**.** *The sample complexity of PAC learning the class REC[$\widetilde{\mathcal{F}_w^\sigma}, T$] of noisy recurrent networks with respect to ramp loss is $\widetilde{O}(w \log(T/\sigma))$.*

One challenge of proving the above theorem is that the analysis involves dealing with *random* hypotheses. Therefore, unlike the usual arguments that bound the covering number of a set of deterministic maps with respect to the ℓ_2 distance, we study the covering number of a class of random maps with respect to the total variation distance. We then invoke some of the recently developed tools in Fathollah Pour and Ashtiani (2022) for bounding these covering numbers. Another challenge is deviating from the usual "unfolding method" and exploiting the fact that in recurrent models a *fixed* function/network is applied recursively.

The mere fact that learning REC[$\widetilde{\mathcal{F}_w^\sigma}, T$] requires less samples compared to its non-noisy counterpart is not entirely unexpected. For classification of long sequences, however, the sample complexity gap is quite drastic (i.e., exponential). We argue that a logarithmic dependency on T is actually more realistic in practical situations: for finite precision machines, one can effectively break the $\Omega(T)$ barrier even for non-noisy networks. To see this, let us choose σ to be a numerically negligible number (e.g., smaller than the numerical precision of our computing device). In this case, the class of noisy and non-noisy networks become effectively the same when implemented on a device with finite numerical precision. But then our upper bound shows a mild logarithmic dependence on $1/\sigma$.

One caveat in the above argument is that the lower bound of Koiran and Sontag (1998) is proved for the 0-1 loss and perhaps not directly comparable to the setting of the upper bound which uses a Lipcshitz surrogate loss. We address this by showing a comparable lower bound in the same setting.

Theorem 2 (Informal version of Theorem 10)**.** *The sample complexity of PAC learning REC[$\mathcal{F}_w, T$] with ramp loss is $\Omega(wT)$.*

In the next section we introduce our notations and define the PAC learning problem. We state the lower bound in Section 3, and the upper bound in Section 5. Sections 6, 7, and 8 provide a high-level proof of our upper bound.

Additional Related Work. Due to space constraints, we postpone the discussion of some additional related work to Appendix A.

2 Preliminaries

2.1 Notations

$\|x\|_1, \|x\|_2$, and $\|x\|_\infty$ denote the ℓ_1, ℓ_2, and ℓ_∞ norms of a vector $x \in \mathbb{R}^d$ respectively. We denote the cardinality of a set S by $|S|$. The set of natural numbers smaller or equal to m is represented by $[m]$. A vector of all zeros is denoted by $0_d = [0 \ \ldots \ 0]^\top \in \mathbb{R}^d$. We use $\mathcal{X} \subseteq \mathbb{R}^d$ as a domain set. We

[1]Ignoring the dependence of the sample complexity on the accuracy and confidence parameters.

will study classes of vector-valued functions; a hypothesis is a Borel function $f : \mathbb{R}^d \to \mathbb{R}^p$, and a hypothesis class $\mathcal{F}$ is a set of such hypotheses.

We find it useful to have an explicit notation—here an overline—for the random versions of the above definitions: $\overline{\mathcal{X}}$ is the set of all random variables defined over $\mathcal{X}$ that admit a generalized density function[2]. $\overline{x} \in \overline{\mathcal{X}}$ is a random variable in this set. To simplify this notation, we sometimes just write $\overline{x} \in \mathbb{R}^d$ rather than $\overline{x} \in \overline{\mathbb{R}^d}$.

$\overline{y} = f(\overline{x})$ is the random variable associated with pushforward of $\overline{x}$ under Borel map $f : \mathbb{R}^d \to \mathbb{R}^p$. We use $\overline{f} : \mathbb{R}^d \to \mathbb{R}^p$ to indicate that the mapping itself is random. Random hypotheses can be applied to both random and non-random inputs—e.g., $\overline{f}(\overline{x})$ and $\overline{f}(x)$[3]. A class of random hypotheses is denoted by $\overline{\mathcal{F}}$.

Definition 3 (Composition of Two Hypothesis Classes)**.** *We denote by $h \circ f$ the function $h(f(x))$ (assuming the range of f and the domain of h are compatible). The composition of two hypothesis classes $\mathcal{F}$ and $\mathcal{H}$ is defined by $\mathcal{H} \circ \mathcal{F} = \{h \circ f \mid h \in \mathcal{H}, f \in \mathcal{F}\}$. Composition of classes of random hypotheses is defined similarly by $\overline{\mathcal{H}} \circ \overline{\mathcal{F}} = \{\overline{h} \circ \overline{f} \mid \overline{h} \in \overline{\mathcal{H}}, \overline{f} \in \overline{\mathcal{F}}\}$.*

2.2 Feedforward neural networks

We will first define some classes associated with feedforward neural networks. Let $\phi(x) = \frac{1}{1+e^{-x}} - \frac{1}{2}$ be the centered sigmoid function. $\Phi : \mathbb{R}^p \to [-1/2, 1/2]^p$ is the element-wise sigmoid activation function defined by $\Phi((x^{(1)}, \dots, x^{(p)})) = (\phi(x^{(1)}), \dots, \phi(x^{(p)}))$.

Definition 4 (Single-Layer Sigmoid Neural Networks)**.** *The class of single-layer sigmoid neural networks with d inputs and p outputs is defined by NET[d, p] $= \{f_W : \mathbb{R}^d \to [-1/2, 1/2]^p \mid f_W(x) = \Phi(W^\top x), W \in \mathbb{R}^{d \times p}\}$.*

Based on Definition 4, we can define the class of multi-layer (feedforward) neural networks (with w weights) as a composition of several single-layer networks. Note that the number of hidden neurons can be arbitrary as long as the total number of weights/parameters is w.

Definition 5 (Multi-Layer Sigmoid Neural Networks)**.** *A class of multi-layer sigmoid networks with p_0 inputs, p_k outputs, and w weights that take inputs in $[-1/2, 1/2]^{p_0}$ is defined by*

$$\textit{MNET}[p_0, p_k, w] = \bigcup \textit{NET}[p_{k-1}, p_k] \circ \ldots \circ \textit{NET}[p_0, p_1]$$

where union is taken over all choices of $(p_1, p_2, \dots, p_{k-1}) \in \mathbb{N}^{k-1}$ that satisfy $\sum_{i=1}^{k} p_i . p_{i-1} = w$. We say MNET[$p_0, p_k, w$] is well-defined if the union is not empty.

Well-definedness basically means that p_0, p_k, and w are compatible. For simplicity, in the above definition we restricted the input domain to $[-1/2, 1/2]^d$. This will help in defining the recurrent versions of these networks (since the input and output domains become compatible). However, our analysis can be easily extended to capture any bounded domain (e.g., $[-B, B]^d$).

2.3 Recursive application of a function and recurrent models

In this section we define REC[$\mathcal{F}, T$] which is the recurrent version of class $\mathcal{F}$ for sequences of length T. Let $v = (a_1, \dots, a_m) \in \mathcal{X}^m$ for $m \in \mathbb{N}$. We define First $(v) = (a_1, \dots, a_{m-1}) \in \mathcal{X}^{m-1}$ and Last $(v) = a_m \in \mathcal{X}$ as functions that return the first $m - 1$ and the last dimensions of the vector v, respectively. Let $u^{(0)}, u^{(1)}, \dots, u^{(T-1)}$ be a sequence of inputs, where $u^{(i)} \in \mathbb{R}^p$, and let $f : \mathbb{R}^s \to \mathbb{R}^q$ be a hypothesis/mapping. In the context of recurrent models, it is useful to define the recurrent application of f on this sequence. Note that out of the q dimensions of the range of f, $q - 1$ of them are recurrent and therefore are fed back to the model. Basically, $f^R(U, t)$ will be the result of applying f on the first t elements of U (with recurrent feedback).

Definition 6 (Recurrent Application of a Function)**.** *Let $U = \left[u^{(0)} \dots u^{(i)} \dots u^{(T-1)}\right] \in \mathbb{R}^{p \times T}$ be a sequence of inputs of length T, where $u^{(i)} \in \mathbb{R}^p$ denotes the i-th column of U for $0 \le i \le T - 1$.*

[2] Both discrete (by using Dirac delta function) and absolutely continuous random variables admit a generalized density function.

[3] Technically, we consider $\overline{f}(x)$ to be $\overline{f}(\overline{\delta_x})$, where $\overline{\delta_x}$ is a random variable with Dirac delta measure on x.

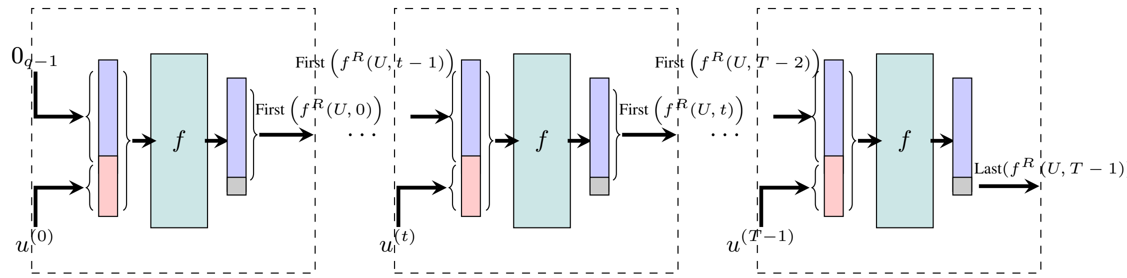

Figure 1: An example of a recurrent model in REC[$\mathcal{F}, T$]. The first $q-1$ dimensions of $f^R(U,t-1)$ is concatenated with $u^{(t)}$ to form the input at time t. The last dimension of $f^R(U,T-1)$ is taken to be the final output of the recurrent model.

Let f be a (random) function from $\mathbb{R}^s$ to $\mathbb{R}^q$, where $s = p+q-1$. Moreover, define $f^R(U,0) = f\left(\begin{bmatrix}0_{q-1} & u^{(0)}\end{bmatrix}^\top\right)$. Then, for any $1 \le t \le T-1$, the recursive application of f is denoted by $f^R : \mathbb{R}^{p\times T} \times [T-1] \to \mathbb{R}^q$ and is defined as $f^R(U,t) = f\left(\begin{bmatrix}\text{First}\left(f^R(U,t-1)\right) & u^{(t)}\end{bmatrix}^\top\right)$.

Now we are ready to define the (recurrent) hypothesis class REC[$\mathcal{F}, T$]. Each hypothesis in this class takes a sequence U of input vectors, and applies a function $f \in \mathcal{F}$ recurrently on the elements of this sequence. The final output will be a real number. We give the formal definition in the following; also see Figure 1 for a visualization.

Definition 7 (Recurrent Class)**.** *Let $s, p, q \in \mathbb{N}$ such that $s = p+q-1$. Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^s$ to $\mathbb{R}^q$. The class of recurrent models with length T that use functions in $\mathcal{F}$ (which we denote by recurring class) as their recurring block is defined by*

$$\text{REC}[\mathcal{F}, T] = \{h : \mathbb{R}^{p\times T} \to \mathbb{R} \mid h(U) = \text{Last}\left(f^R(U,T-1)\right), f \in \mathcal{F}\}$$

For example, REC[MNET[p_0, p_k, w], T] is the class of (real-valued) recurrent neural networks with length T that use MNET[p_0, p_k, w] as their recurring block. We say that REC[MNET[p_0, p_k, w], T] is *well-defined* if MNET[p_0, p_k, w] is well-defined and also the input/output dimensions are compatible (i.e., $p_0 \ge p_k$).

2.4 PAC learning with ramp loss

In this section we formulate the PAC learning model for classification with respect to the ramp loss. The use of ramp loss is natural for classification (see e.g., Boucheron et al. (2005); Bartlett et al. (2006)) and the main features of the ramp loss that we are going to exploit are boundedness and Lipschitzness. We start by introducing the ramp loss.

Definition 8 (Ramp Loss)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be a hypothesis and let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. Let $(x,y) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{-1, 1\}$. The ramp loss of f with respect to margin parameter $\gamma > 0$ is defined as $l_\gamma(f,x,y) = r_\gamma(-f(x).y)$, where r_γ is the ramp function defined by*

$$r_\gamma(x) = \begin{cases} 0 & x < -\gamma, \\ 1 + \frac{x}{\gamma} & -\gamma \le x \le 0 \\ 1 & x \ge 0. \end{cases}$$

Definition 9 (Agnostic PAC Learning with Respect to Ramp Loss)**.** *We say that a hypothesis class $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ is agnostic PAC learnable with respect to ramp loss with margin parameter $\gamma > 0$ if there exists a learner $\mathcal{A}$ and a function $m : (0,1)^2 \to \mathbb{N}$ with the following property: For every distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, 1\}$ and every $\epsilon, \delta \in (0,1)$, if S is a set of $m(\epsilon, \delta)$ i.i.d. samples from $\mathcal{D}$, then with probability at least $1-\delta$ (over the randomness of S) we have*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma\left(\mathcal{A}(S), x, y\right)\right] \le \inf_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma\left(\mathcal{A}(S), x, y\right)\right] + \epsilon.$$

The *sample complexity* of PAC learning $\mathcal{F}$ with respect to ramp loss is denoted by $m_{\mathcal{F}}(\epsilon, \delta)$, which is the minimum number of samples required for learning $\mathcal{F}$ (among all learners $\mathcal{A}$). The definition of agnostic PAC learning with respect to ramp loss works for any value of γ and when we are analyzing the sample complexity we consider it to be a fixed constant.

3 A lower bound for sample complexity of learning recurrent neural networks

In this section, we consider the sample complexity of PAC learning sigmoid recurrent neural networks with respect to ramp loss. Particularly, we state a lower bound on the sample complexity of the class REC[MNET[p_0, p_k, w], T] of all sigmoid recurrent neural networks with length T that use multi-layer neural networks with w weights as their recurring block. The main message is that this sample complexity grows at least linearly with T.

Theorem 10 (Sample Complexity Lower Bound for Recurrent Neural Networks)**.** *For every $T \geq 3$ and $w \geq 19$ there exists a well-defined class $\mathcal{H}_w$ = REC[MNET[p_0, p_k, w], T] and a universal constant $C > 0$ such that for every $\epsilon, \delta \in (0, 1/40)$ we have*

$$m_{\mathcal{H}_w}(\epsilon, \delta) \geq C. \left(\frac{wT + \log(1/\delta)}{\epsilon^2} \right).$$

The proof of the above lower bound is based on a similar result due to Sontag et al. (1998). However, the argument in Sontag et al. (1998) is for PAC learning with respect to 0-1 loss. To extend this result for the ramp loss, we construct a binary-valued class $\mathcal{F}_w = \{f : f(U) = \text{sign}(h(U)), h \in \mathcal{H}_w\}$ where sign $(x) = 1$ if $x \geq 0$ and sign $(x) = -1$ if $x < 0$. We prove that every function $f \in \mathcal{F}_w$ can be related to another function $h \in \mathcal{H}_w$ such that the ramp loss of h is almost equal to the zero-one loss of f. This is formalized in the following lemma, which is a key result in proving Theorem 10. The proof of Theorem 10 and Lemma 11 can be found in Appendix C.

Lemma 11. *Let $\mathcal{H}_w$ = REC[MNET[p_0, p_k, w], T] be a well-defined class and let $\mathcal{F}_w = \{f : [-1/2, 1/2]^{p \times T} \to \{-1, 1\} \mid f(U) = sign\,(h(U)), h \in \mathcal{H}_w\}$. Then, for every distribution $\mathcal{D}$ over $[-1/2, 1/2]^{p \times T} \times \{-1, 1\}$, $\eta > 0$, and every function $f \in \mathcal{F}_w$ there exists a function $h \in \mathcal{H}_w$ such that $\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma\left(h, U, y\right)\right] \leq \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l^{0-1}\left(f, U, y\right)\right] + \eta$ where $l^{0-1}(f, U, y) = 1\left\{f(U) \neq y\right\}$.*

4 Noisy recurrent neural networks

In this section, we will define classes of noisy recurrent neural networks. Let us first define the singleton Gaussian noise class, which contains a single additive Gaussian noise function.

Definition 12 (The Gaussian Noise Class)**.** *The d-dimensional noise class with scale $\sigma \geq 0$ is denoted by $\overline{\mathcal{G}_{\sigma,d}} = \{\overline{g_{\sigma,d}}\}$. Here, $\overline{g_{\sigma,d}} : \mathbb{R}^d \to \mathbb{R}^d$ is a random function defined by $\overline{g_{\sigma,d}}(\overline{x}) = \overline{x} + \overline{z}$, where $\overline{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$. When it is clear from the context we drop d and write $\overline{\mathcal{G}_\sigma} = \{\overline{g_\sigma}\}$.*

The following is the noisy version of multi-layer networks in Definition 5. Basically, Gaussian noise is composed (Definition 3) before each layer.

Definition 13 (Noisy Multi-Layer Sigmoid Neural Networks)**.** *The class of all noisy multi-layer sigmoid networks with w weights that take values in $[-1/2, 1/2]^{p_0}$ as input and output values in $[-1/2, 1/2]^{p_k}$ is defined by*

$$\overline{\textit{MNET}_\sigma}[p_0, p_k, w] = \bigcup \textit{NET}[p_{k-1}, p_k] \circ \ldots \circ \overline{\mathcal{G}_\sigma} \circ \textit{NET}[p_1, p_2] \circ \overline{\mathcal{G}_\sigma} \circ \textit{NET}[p_0, p_1] \circ \overline{\mathcal{G}_\sigma},$$

where $\sigma \geq 0$ is scale of the Gaussian noise and the union is taken over all choices of $(p_1, p_2, \ldots, p_{k-1}) \in \mathbb{N}^{k-1}$ that satisfy $\sum_{i=1}^{k} p_i.p_{i-1} = w$.

Similar to the deterministic case, $\overline{\text{MNET}_\sigma}[p_0, p_k, w]$ is said to be well-defined if the union is not empty (i.e., p_0, p_k and w are compatible). We can use Definition 7 to create recurrent versions of the above class. For example, REC[$\overline{\text{MNET}_\sigma}[p_0, p_k, w], T$] is a class of recurrent (and random) hypotheses for sequence of length T that use $\overline{\text{MNET}_\sigma}[p_0, p_k, w]$ as their recurring block. Again, similar to the deterministic case, we say REC[$\overline{\text{MNET}_\sigma}[p_0, p_k, w], T$] is well-defined if p_0, p_k and w are compatible and $\overline{\text{MNET}_\sigma}[p_0, p_k, w]$ is well-defined.

5 PAC learning noisy recurrent neural networks

In section 3, we established an $\Omega(T)$ lower bound on the sample complexity of learning recurrent networks (i.e., REC[MNET$[p_0, p_k, w], T$]). In this section, we consider a related class (based on noisy recurrent neural networks) and show that the dependence of sample complexity on T is only $O(\log T)$. In particular, $\overline{\mathcal{G}_\sigma} \circ$ REC[$\overline{\text{MNET}_\sigma}[p_0, p_k, w], T$] can be regarded as a (noisy) sibling of REC[MNET$[p_0, p_k, w], T$]. Since it is more standard to define PAC learnability for deterministic hypotheses, we define the deterministic version of the above class by derandomization[4].

Definition 14 (Derandomization by Expectation). *Let $\mathcal{F}$ be a class of (random) functions from $\mathbb{R}^{p\times T}$ to $\mathbb{R}^q$. The derandomization of a function class $\overline{\mathcal{F}}$ by expectation is defined as $\mathcal{E}(\overline{\mathcal{F}}) = \{h : \mathbb{R}^{p\times T} \to \mathbb{R}^q \mid h(u) = \mathbb{E}_{\overline{f}}\left[\,\overline{f}(u)\right], \overline{f} \in \overline{\mathcal{F}}\}$.*

We show that, contrary to Theorem 10, the sample complexity of PAC learning the (derandomized) class of noisy recurrent neural networks, $\mathcal{E}(\overline{\mathcal{G}_\sigma} \circ \text{REC}[\overline{\text{MNET}_\sigma}[p_0, p_k, w], T])$, grows at most logarithmically with T while it still enjoys the same linear dependence on w. This is formalized in the following theorem (see Appendix D for a proof).

Theorem 15 (Main Result). *Let $\overline{\mathcal{Q}_w} = \overline{\mathcal{G}_\sigma} \circ$ REC[$\overline{\text{MNET}_\sigma}[p_0, p_k, w], T$] be any well-defined class and assume $T \in \mathbb{N}, 0 < \sigma < 1, \epsilon, \delta \in (0, 1)$. Then the sample complexity of learning $\mathcal{H}_w = \mathcal{E}(\overline{\mathcal{Q}_w})$ is upper bounded by*

$$m_{\mathcal{H}_w}(\epsilon, \delta) = O\left(\frac{w \log\left(\frac{wT}{\epsilon\sigma} \log\left(\frac{wT}{\epsilon\sigma}\right)\right) + \log\left(1/\delta\right)}{\epsilon^2}\right) = \widetilde{O}\left(\frac{w \log\left(\frac{T}{\sigma}\right) + \log(1/\delta)}{\epsilon^2}\right),$$

where $\widetilde{O}$ hides logarithmic factors.

One feature of the above theorem is the mild logarithmic dependence on $1/\sigma$. Therefore, we can take σ to be numerically negligible and still get a significantly smaller sample complexity compared to the deterministic case for large T. Note that adding such small values of noise would not change the empirical outcome of RNNs on finite precision computers.

The milder (logarithmic) dependency on T is achieved by a novel analysis that involves bounding the covering number of noisy recurrent networks with respect to the total variation distance. Also, instead of "unfolding" the network, we exploit the fact that the same function/hypothesis is being used recurrently. We also want to emphasize that the above bound does not depend on the norms of weights of the network. Achieving this is challenging, since a little bit of noise in a previous layer can change the output of the next layer drastically. The next few sections are dedicated to give a high-level proof of this theorem.

6 Covering numbers: the classical view

One of the main tools to derive sample complexity bounds for learning a class of functions is studying their covering numbers. In this section we formalize this classic tool.

Definition 16 (Covering Number). *Let $(\mathcal{X}, \rho)$ be a metric space. A set $A \subset \mathcal{X}$ is ϵ-covered by a set $C \subseteq A$ with respect to ρ, if for all $a \in A$ there exists $c \in C$ such that $\rho(a, c) \leq \epsilon$. We denote by $N(\epsilon, A, \rho)$ the cardinality of the smallest set C that ϵ-covers A and we refer to is as the ϵ-covering number of A with respect to metric ρ.*

The notion of covering number is defined with respect to a metric ρ. We now give the definition of extended metrics, which we will use to define *uniform* covering numbers. The extended metrics can be seen as measures of distance between two hypotheses on a given input set.

Definition 17 (Extended Metrics). *Let $(\mathcal{X}, \rho)$ be a metric space. Let $u = (a_1, \ldots, a_m), v = (b_1, \ldots, b_m) \in \mathcal{X}^m$ for $m \in \mathbb{N}$. The ∞-extended and ℓ_2-extended metrics over $\mathcal{X}^m$ are defined by $\rho^{\infty,m}(u, v) = \sup_{1\leq i\leq m} \rho(a_i, b_i)$ and $\rho^{\ell_2,m}(u, v) = \sqrt{\frac{1}{m}\sum_{i=1}^m (\rho(a_i, b_i))^2}$, respectively. We drop m and use ρ^∞ or ρ^{ℓ_2} if it is clear from the context.*

[4] One can also define PAC learnability for a class of random hypotheses and get a similar result without taking the expectation. However, working with a deterministic class helps to contrast the result with that of Theorem 10.

A useful property about extended metrics is that the ∞-extended metric always upper bounds the ℓ_2-extended metric, i.e., $\rho^{\ell_2}(u,v) \le \rho^{\infty}(u,v)$ for all $u, v \in \mathcal{X}$. Based on the above definition of extended metrics, we define the uniform covering number of a hypothesis class with respect to $\|.\|_2$.

Definition 18 (Uniform Covering Number with Respect to $\|.\|_2$)**.** *Let $\mathcal{F}$ be a hypothesis class of functions from $\mathcal{X}$ to $\mathcal{Y}$. For a set of inputs $S = \{x_1, x_2, \ldots, x_m\} \subseteq \mathcal{X}$, we define the restriction of $\mathcal{F}$ to S as $\mathcal{F}_{|S} = \{(f(x_1), f(x_2), \ldots, f(x_m)) : f \in \mathcal{F}\} \subseteq \mathcal{Y}^m$. The uniform ϵ-covering numbers of hypothesis class $\mathcal{F}$ with respect to $\|.\|_2^{\infty}$, $\|.\|_2^{\ell_2}$ are denoted by $N_U(\epsilon, \mathcal{F}, m, \|.\|_2^{\infty})$ and $N_U(\epsilon, \mathcal{F}, m, \|.\|_2^{\ell_2})$ and are the maximum values of $N(\epsilon, \mathcal{F}_{|S}, \|.\|_2^{\infty,m})$ and $N(\epsilon, \mathcal{F}_{|S}, \|.\|_2^{\ell_2,m})$ over all $S \subseteq \mathcal{X}$ with $|S| = m$, respectively.*

The following theorem connects the notion of uniform covering number with PAC learning. It converts a bound on the $\|.\|_2^{\ell_2}$ uniform covering number of a hypothesis class to a bound on the sample complexity of PAC learning the class; see Appendix E for a more detailed discussion.

Theorem 19. *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. Then there exists an algorithm $\mathcal{A}$ with the following property: For every distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, 1\}$ and every $\epsilon, \delta \in (0,1)$, if S is a set of m i.i.d. samples from $\mathcal{D}$, then with probability at least $1-\delta$ (over the randomness of S),*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma\left(\mathcal{A}(S), x, y\right)\right]$$
$$\le \inf_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma\left(f, x, y\right)\right] + 16\epsilon + \frac{24}{\sqrt{m}}\sqrt{\ln N_U(\gamma\epsilon, \mathcal{F}, m, \|.\|_2^{\ell_2})} + 6\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

Moreover, the algorithm that returns the function with the minimum error on S satisfies the above property (i.e., Algorithm $\mathcal{A}$ such that $\mathcal{A}(S) = \arg\min_{f\in\mathcal{F}} \frac{1}{|S|}\sum_{(x,y)\in S} l_\gamma\left(f, x, y\right)$).

7 Total variation covers for random hypotheses

One idea to prove a generalization bound for noisy neural networks is to bound their covering numbers. However, noisy neural networks are random functions, and therefore their behaviours on a sample set cannot be directly compared. Instead, one can compare the output distributions of a random function on two sample sets. We therefore use the recently developed tools from Fathollah Pour and Ashtiani (2022) to define and study covering numbers for random hypotheses. These covering numbers are defined based on metrics between distributions. Specifically, our analysis is based on the notion of uniform covering number with respect to total variation distance.

Definition 20 (Total Variation Distance)**.** *Let μ and ν denote two probability measures over $\mathcal{X}$ and let Ω be the Borel sigma-algebra over $\mathcal{X}$. The TV distance between μ and ν is defined by*

$$d_{TV}(\mu, \nu) = \sup_{B\in\Omega} |\mu(B) - \nu(B)|.$$

Furthermore, if μ and ν have densities f and g then

$$d_{TV}(\mu,\nu) = \sup_{B\in\Omega}\Big|\int_B (f(x)-g(x))dx\Big| = \frac{1}{2}\int_{\mathcal{X}} |f(x)-g(x)|\,dx = \frac{1}{2}\|f-g\|_1.$$

For two random variables $\overline{x}$ and $\overline{y}$ with probability measures μ and ν we sometimes abuse the notation and write $d_{TV}(\overline{x}, \overline{y})$ instead of $d_{TV}(\mu,\nu)$. For example, we write $d_{TV}(\overline{f_1}(\overline{x}), \overline{f_2}(\overline{x}))$ in order to refer to the Total Variation (TV) distance between pushforwards of $\overline{x}$ under mappings $\overline{f_1}$ and $\overline{f_2}$. We also write $d_{TV}^{\infty,m}\left(\left(\overline{f_1}(\overline{x_1}), \ldots, \overline{f_1}(\overline{x_m})\right), \left(\overline{f_2}(\overline{x_1}), \ldots, \overline{f_2}(\overline{x_m})\right)\right)$ to refer to the extended TV distance between mappings of the set $S = \{\overline{x_1}, \ldots, \overline{x_m}\}$ by $\overline{f_1}$ and $\overline{f_2}$. We use the extended total variation distance to define the uniform covering number for classes of random hypotheses.

Definition 21 (Uniform Covering Number for Classes of Random Hypotheses)**.** *Let $\overline{\mathcal{F}}$ be a class of random hypotheses from $\overline{\mathcal{X}}$ to $\overline{\mathcal{Y}}$. For a set of random variables $\overline{S} = \{\overline{x_1}, \overline{x_2}, \ldots, \overline{x_m}\} \subseteq \overline{\mathcal{X}}$, the restriction of $\overline{\mathcal{F}}$ to $\overline{S}$ is defined as $\overline{\mathcal{F}}_{|\overline{S}} = \{(\overline{f}(\overline{x_1}), \overline{f}(\overline{x_2}), \ldots, \overline{f}(\overline{x_m})) : \overline{f} \in \overline{\mathcal{F}}\} \subseteq \overline{\mathcal{Y}}^m$. Let $\Gamma \subseteq \overline{\mathcal{X}}$. The uniform ϵ-covering numbers of $\overline{\mathcal{F}}$ with respect to Γ and d_{TV}^{∞} is defined by*

$$N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\infty}, \Gamma) = \sup_{S\subseteq\Gamma, |S|=m} N(\epsilon, \overline{\mathcal{F}}_{|\overline{S}}, d_{TV}^{\infty,m}).$$

Some hypothesis classes that we analyze (e.g., single-layer noisy neural networks) may have "global" total variation covers that do not depend on m. This will be addressed with the following notation: $N_U(\epsilon, \overline{\mathcal{F}}, \infty, \rho^\infty, \Gamma) = \lim_{m\to\infty} N_U(\epsilon, \overline{\mathcal{F}}, m, \rho^\infty, \Gamma)$. The set Γ in Definition 21 is used to define the input domain for which we want to find the covering number of a class of random hypotheses. For instance, some of the covers that we see are derived with respect to inputs with bounded domain or some need the input to be first smoothed by Gaussian noise. In this paper, we will be working with the following choices of Γ

- $\Gamma = \overline{\mathcal{X}_d}$ and $\Gamma = \overline{\mathcal{X}_{B,d}}$: the set of all random variables defined over $\mathbb{R}^d$ and $[-B, B]^d$, respectively, that admit a generalized density function. For example, we use $\overline{\mathcal{X}_{0.5,d}}$ to address the set of random variables in $[-1/2, 1/2]^d$.
- $\Gamma = \overline{\Delta_{p\times T}} = \{\overline{U} \mid \overline{U} = \begin{bmatrix}\overline{\delta_{u^{(0)}}} & \dots & \overline{\delta_{u^{(T-1)}}}\end{bmatrix}^\top, u^{(i)} \in \mathbb{R}^p\}$ and $\Gamma = \overline{\Delta_{B,p\times T}} = \{\overline{U} \mid \overline{U} = \begin{bmatrix}\overline{\delta_{u^{(0)}}} & \dots & \overline{\delta_{u^{(T-1)}}}\end{bmatrix}^\top, u^{(i)} \in [-B, B]^p\}$, where $\overline{\delta_{u^{(i)}}}$ is the random variable associated with Dirac delta measure on $u^{(i)}$. Note that $\overline{\Delta_{B,p\times T}} \subset \overline{\Delta_{p\times T}}$.
- $\Gamma = \overline{\mathcal{G}_{\sigma,d}} \circ \overline{\mathcal{X}_{B,d}} = \{\overline{g_{\sigma,d}}(\overline{x}) \mid \overline{x} \in \overline{\mathcal{X}_{B,d}}\}$: all members of $\overline{\mathcal{X}_{B,d}}$ after being "smoothed" by adding (convolving the density with) Gaussian noise.

We mentioned in Section 6 that a bound on the $\|.\|_2^{\ell_2}$ uniform covering number can be connected to a bound on sample complexity of PAC learning. We now show that a bound on d_{TV}^∞ covering number of a class of random hypotheses can be turned into a bound on the $\|.\|_2^{\ell_2}$ covering number of its derandomized version and, thus, PAC learning it.

Theorem 22 ($\|.\|_2^{\ell_2}$ Cover of $\mathcal{E}(\mathcal{F})$ From d_{TV}^∞ Cover of $\mathcal{F}$ (Fathollah Pour and Ashtiani, 2022))**.** *Let $\overline{\mathcal{F}}$ be a class of functions from $\mathbb{R}^{p\times T}$ to $[-B, B]^q$. Then for every $\epsilon > 0$ and $m \in \mathbb{N}$ we have*

$$N_U(2B\epsilon\sqrt{q}, \mathcal{E}(\overline{\mathcal{F}}), m, \|.\|_2^{\ell_2}) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^\infty, \overline{\Delta_{p\times T}}) \leq N_U(\epsilon, \overline{\mathcal{F}}, \infty, d_{TV}^\infty, \overline{\Delta_{p\times T}}).$$

8 Bounding the covering number of recurrent models

In Section 6, we mentioned that finding a bound on covering number of a hypothesis class is a standard approach to bound its sample complexity. In the previous section, we introduced a new notion of covering number with respect to total variation distance that was developed by Fathollah Pour and Ashtiani (2022). We showed how this notion can be related to PAC learning for classes of random hypotheses. In the following, we give an overview of the techniques used to find a bound on the d_{TV}^∞ covering number of the class of noisy recurrent models. We also discuss why this bound results in a sample complexity that has a milder logarithmic dependency on T, compared to bounds proved by "unfolding" the recurrence and replacing the recurrent model with the T-fold composition.

One advantage of analyzing the uniform covering number with respect to TV distance is that it comes with a useful composition tool. The following theorem basically states that when two classes of hypotheses have bounded TV covers, their composition class has a bounded cover too. Note that such a result does not hold for the usual definition of covering number (e.g., Definition 18); see Fathollah Pour and Ashtiani (2022) for details.

Theorem 23 (TV Cover for Composition of Random Classes, Lemma 18 of Fathollah Pour and Ashtiani (2022))**.** *Let $\overline{\mathcal{F}}$ be a class of random hypotheses from $\mathbb{R}^d$ to $\mathbb{R}^p$ and $\overline{\mathcal{H}}$ be a class of random hypotheses from $\mathbb{R}^p$ to $\mathbb{R}^q$. For any $\epsilon_1, \epsilon_2 > 0$ and $m \in \mathbb{N}$, denote $N_1 = N_U\left(\epsilon_1, \overline{\mathcal{F}}, m, d_{TV}^\infty, \overline{\mathcal{X}_d}\right)$. Then we have,*

$$N_U\left(\epsilon_1 + \epsilon_2, \overline{\mathcal{H}} \circ \overline{\mathcal{F}}, m, d_{TV}^\infty, \overline{\mathcal{X}_d}\right) \leq N_U\left(\epsilon_2, \overline{\mathcal{H}}, mN_1, d_{TV}^\infty, \overline{\mathcal{X}_p}\right).N_1.$$

An approach to bound the TV uniform covering number of a recurrent model REC$[\overline{\mathcal{F}}, T]$ is to consider it as the T-fold composition $\overline{\mathcal{F}} \circ \overline{\mathcal{F}} \dots \circ \overline{\mathcal{F}}$. One can then use a similar analysis to that of Fathollah Pour and Ashtiani (2022) to bound the covering number of the T-fold composition. Unfortunately, this approach fails to capture the fact that a *fixed* function $\overline{f} \in \overline{\mathcal{F}}$ is applied recursively, and therefore results in a sample complexity bound that grows at least linearly with T.

Instead, we take another approach to bound the covering number of recurrent models. Intuitively, we notice that any function in the T-fold composite class $\overline{\mathcal{F}} \circ \dots \circ \overline{\mathcal{F}} = \{\overline{f_1} \circ \dots \circ \overline{f_T} \mid \overline{f_1}, \dots, \overline{f_T} \in$

$\overline{\mathcal{F}}\}$ is determined by T functions from $\overline{\mathcal{F}}$. On the other hand, any function in REC$[\overline{\mathcal{F}}, T] = \left\{\overline{h} \mid \overline{h}(U) = \text{Last}\left(\overline{f}^R(U, T-1)\right)\right\}$ is only defined by one function in $\overline{\mathcal{F}}$ and the capacity of this class must not be as large as the capacity of $\overline{\mathcal{F}} \circ \ldots \circ \overline{\mathcal{F}}$. Interestingly, data processing inequality for total variation distance (Lemma 27) suggests that if two functions $\overline{f}$ and $\overline{\hat{f}}$ are "globally" close to each other with respect to TV distance (i.e., $d_{TV}(\overline{f}(\overline{x}), \overline{\hat{f}}(\overline{x})) \leq \epsilon$ for every $\overline{x}$ in the domain), then $d_{TV}(\overline{f}(\overline{f}(\overline{x})), \overline{\hat{f}}(\overline{\hat{f}}(\overline{x}))) \leq 2\epsilon$ (i.e., $\overline{f} \circ \overline{f}$ and $\overline{\hat{f}} \circ \overline{\hat{f}}$ are also close to each other). By applying the data processing inequality recursively, we can see that for the T-fold composition we have $d_{TV}(\overline{f} \circ \ldots \circ \overline{f}(\overline{x}), \overline{\hat{f}} \circ \ldots \circ \overline{\hat{f}}(\overline{x})) \leq \epsilon T$. The above approach results in the following theorem which bounds the ϵ-covering number of a noisy recurrent model with respect to TV distance by the (ϵ/T)-covering number of its recurring class. Intuitively, this theorem helps us to bound the covering number of noisy recurrent models using the bounds obtained for their non-recurrent versions. Here, Gaussian noise is added to both the input of the model (i.e., $\overline{\mathcal{F}_\sigma} = \overline{\mathcal{F}} \circ \overline{\mathcal{G}_\sigma}$) and the output of the model (by composing with $\overline{\mathcal{G}_\sigma}$).

Theorem 24 (TV Covering Number of $\overline{\mathcal{G}_\sigma} \circ$ REC$[\overline{\mathcal{F}_\sigma}, T]$ From $\overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{F}_\sigma}$)**.** *Let $s, p, q \in \mathbb{N}$ such that $s = p + q - 1$. Let $\overline{\mathcal{F}}$ be a class of functions from $\overline{\mathcal{X}_{B,s}}$ to $\overline{\mathcal{X}_{B,q}}$ and denote by $\overline{\mathcal{F}_\sigma} = \overline{\mathcal{F}} \circ \overline{\mathcal{G}_{\sigma,s}}$ the class of its composition with noise. Then we have*

$$N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ REC[\overline{\mathcal{F}_\sigma}, T], \infty, d_{TV}^\infty, \overline{\Delta_{B,p\times T}}\right) \leq N_U\left(\epsilon/T, \overline{\mathcal{G}_{\sigma,q}} \circ \overline{\mathcal{F}_\sigma}, \infty, d_{TV}^\infty, \overline{\mathcal{X}_{B,s}}\right).$$

For using this theorem, one needs to have a finer ϵ/T-cover for the recurring class. As we will see in the next section, this will translate into a mild logarithmic sample complexity dependence on T.

8.1 Covering noisy recurrent networks

An example of $\overline{\mathcal{F}_\sigma}$ is the class $\overline{\text{MNET}_\sigma}[p_0, p_k, w]$ of well-defined noisy multi-layer networks (Definition 13). Theorem 24 suggests that a bound on the covering number of $\overline{\mathcal{G}_\sigma} \circ$ REC$[\overline{\text{MNET}_\sigma}[p_0, p_k, w], T]$ can be found from a bound for $\overline{\mathcal{G}_\sigma} \circ \overline{\text{MNET}_\sigma}[p_0, p_k, w]$. We use the following theorem as a bound for the class of single-layer noisy sigmoid networks together with theorem 23 to bound the covering number of $\overline{\mathcal{G}_\sigma} \circ \overline{\text{MNET}_\sigma}[p_0, p_k, w]$ (see Appendix D, Theorem 38).

Theorem 25 (A TV Cover for Single-Layer Noisy Neural Networks, Theorem 25 of Fathollah Pour and Ashtiani (2022))**.** *For every $p, d \in \mathbb{N}, \epsilon > 0, \sigma < 5d/\epsilon$ we have*

$$\log N_U(\epsilon, \overline{\mathcal{G}_{\sigma,p}} \circ NET[d,p], \infty, d_{TV}^\infty, \overline{\mathcal{G}_{\sigma,d}} \circ \overline{\mathcal{X}_{0.5,d}}) \leq p(d+1)\log\left(30\frac{d^{5/2}\sqrt{\ln\left(\frac{5d-\epsilon\sigma}{\epsilon\sigma}\right)}}{\epsilon^{3/2}\sigma^2}\ln\left(\frac{5d}{\epsilon\sigma}\right)\right).$$

Interestingly, the above bound (on the logarithm of the covering number) is logarithmic with respect to $1/\epsilon$. We will extend this result to multi-layer noisy networks, and then apply Theorem 24 to obtain the following bound on the covering number noisy recurrent neural networks. Crucially, the dependency (of the logarithm of the covering number) on T is only logarithmic.

Theorem 26 (A TV Covering Number Bound for Noisy Sigmoid Recurrent Networks)**.** *Let $T \in \mathbb{N}$. For every $\epsilon, \sigma \in (0,1)$ and every well-defined class REC$[\overline{MNET_\sigma}[p_0, p_k, w], T]$ we have*

$$\log N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ REC[\overline{MNET_\sigma}[p_0, p_k, w], T], \infty, d_{TV}^\infty, \overline{\Delta_{0.5,p\times T}}\right)$$
$$= O\left(w \log\left(\frac{wT}{\epsilon\sigma}\log\left(\frac{wT}{\epsilon\sigma}\right)\right)\right) = \widetilde{O}\left(w\log\left(\frac{T}{\epsilon\sigma}\right)\right).$$

Finally, we turn the above bound into a $\|.\|_2^{\ell_2}$ covering number bound for the derandomized function $\mathcal{E}\left(\overline{\mathcal{G}_\sigma} \circ \text{REC}[\overline{\text{MNET}_\sigma}[p_0, p_k, w], T]\right)$ by an application of Theorem 22. We then upper bound the sample complexity by the logarithm of covering number (see Theorem 19) and conclude Theorem 15.

Limitations and future work. Our results are derived for sigmoid (basically bounded, monotone, and Lipschitz) activation functions. It is open whether such results can be proved for unbounded activation functions such as RELU. Our results are theoretical and we leave empirical evaluations on the performance of noisy networks to future work.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Can sgd learn recurrent neural networks with provable generalization? *Advances in Neural Information Processing Systems*, 32, 2019.

Martin Anthony, Peter L Bartlett, Peter L Bartlett, et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Peter Bartlett. For valid generalization the size of the weights is more important than the size of the network. *Advances in neural information processing systems*, 9, 1996.

Peter Bartlett, Vitaly Maiorov, and Ron Meir. Almost linear vc dimension bounds for piecewise polynomial networks. *Advances in neural information processing systems*, 11, 1998.

Peter L Bartlett and Wolfgang Maass. Vapnik-chervonenkis dimension of neural nets. *The handbook of brain theory and neural networks*, pages 1188–1192, 2003.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Eric Baum and David Haussler. What size net gives valid generalization? *Advances in neural information processing systems*, 1, 1988.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31, 2018.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Minshuo Chen, Xingguo Li, and Tuo Zhao. On generalization bounds of a family of recurrent neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1233–1243. PMLR, 2020.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

Richard M Dudley. Universal donsker classes and metric entropy. In *Selected Works of RM Dudley*, pages 345–365. Springer, 2010.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Alireza Fathollah Pour and Hassan Ashtiani. Benefits of additive noise in composing classes with bounded capacity. *Advances in Neural Information Processing Systems*, 35:32709–32722, 2022.

Wei Gao and Zhi-Hua Zhou. Dropout rademacher complexity of deep neural networks. *Science China Information Sciences*, 59(7):1–12, 2016.

Paul W Goldberg and Mark R Jerrum. Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18(2):131–148, 1995.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30, 2017.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 33:9925–9935, 2020.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19:1–44, 2018.

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136. PMLR, 2020.

Kam-Chuen Jim, C Lee Giles, and Bill G Horne. An analysis of noise in recurrent neural networks: convergence and generalization. *IEEE Transactions on neural networks*, 7(6):1424–1438, 1996.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

Pascal Koiran and Eduardo D Sontag. Vapnik-chervonenkis dimension of recurrent neural networks. *Discrete Applied Mathematics*, 86(1):63–79, 1998.

Soon Hoe Lim, N Benjamin Erichson, Liam Hodgkinson, and Michael W Mahoney. Noisy recurrent neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Philip M Long and Hanie Sedghi. Size-free generalization bounds for convolutional neural networks. In *International Conference on Learning Representations*, 2020.

Wolfgang Maass. Neural nets with superlinear vc-dimension. *Neural Computation*, 6(5):877–884, 1994.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Vaishnavh Nagarajan and Zico Kolter. Deterministic pac-bayesian generalization bounds for deep networks via generalizing noise-resilience. In *International Conference on Learning Representations*, 2018.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. *Advances in Neural Information Processing Systems*, 32, 2019.

Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, pages 3526–3545. PMLR, 2021.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401. PMLR, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Yao Qin, Dongjin Song, Haifeng Cheng, Wei Cheng, Guofei Jiang, and Garrison W Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2627–2633, 2017.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240. PMLR, 2016.

Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Eduardo D Sontag et al. Vc dimension of neural networks. *NATO ASI Series F Computer and Systems Sciences*, 168:69–96, 1998.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pages 3437–3452. PMLR, 2020.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Zhuozhuo Tu, Fengxiang He, and Dacheng Tao. Understanding generalization in recurrent neural networks. In *International Conference on Learning Representations*, 2020.

Mathukumalli Vidyasagar. *A theory of learning and generalization: with applications to neural networks and control systems*. Springer-Verlag, 1997.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066. PMLR, 2013.

Haotian Wang, Wenjing Yang, Zhenyu Zhao, Tingjin Luo, Ji Wang, and Yuhua Tang. Rademacher dropout: An adaptive dropout for deep neural network via optimizing generalization gap. *Neurocomputing*, 357:177–187, 2019.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Jiong Zhang, Qi Lei, and Inderjit Dhillon. Stabilizing gradients for deep neural networks via efficient svd parameterization. In *International Conference on Machine Learning*, pages 5806–5814. PMLR, 2018.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

Jingyu Zhao, Feiqing Huang, Jia Lv, Yanjie Duan, Zhen Qin, Guodong Li, and Guangjian Tian. Do rnn and lstm have long memory? In *International Conference on Machine Learning*, pages 11365–11375. PMLR, 2020.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach. In *International Conference on Learning Representations (ICLR)*, 2019.

A More on related work

There is plethora of work on generalization in neural networks. There are a family of approaches that aim to bound the VC-dimension of neural networks. (Baum and Haussler, 1988; Maass, 1994; Goldberg and Jerrum, 1995; Vidyasagar, 1997; Sontag et al., 1998; Koiran and Sontag, 1998; Bartlett et al., 1998; Bartlett and Maass, 2003; Bartlett et al., 2019). These approaches result in generalization bounds that are dependent on the number of parameters. Another family of approaches are aimed at obtaining generalization bounds that are dependent on the norms of the weights and Lipschitz continuity properties of the network (Bartlett, 1996; Anthony et al., 1999; Zhang, 2002; Neyshabur et al., 2015; Bartlett et al., 2017; Neyshabur et al., 2018; Golowich et al., 2018; Arora et al., 2018; Nagarajan and Kolter, 2018; Long and Sedghi, 2020). It has been observed that these generalization bounds are usually vacuous in practice. One speculation is that the implicit bias of gradient descent (Gunasekar et al., 2017; Arora et al., 2019; Ji et al., 2020; Chizat and Bach, 2020; Ji and Telgarsky, 2021) can lead to benign overfitting (Belkin et al., 2018, 2019; Bartlett et al., 2020, 2021). It has also been conjectured that uniform convergence theory may not be able to fully capture the performance of neural networks in practice (Nagarajan and Kolter, 2019; Zhang et al., 2021). It has been shown that there are data-dependent approaches that can achieve non-vacuouys bounds(Dziugaite and Roy, 2017; Zhou et al., 2019; Negrea et al., 2019). There are also other approaches that are independent of data (Arora et al., 2018); see Fathollah Pour and Ashtiani (2022) for more details.

Adding different types of noise such as dropout noise (Srivastava et al., 2014), DropConnect (Wan et al., 2013), and Denoising AutoEncoders (Vincent et al., 2008) are shown to be helpful in training neural networks. Wang et al. (2019) and Gao and Zhou (2016) theoretically analyze the generalization under dropout noise. More recently, Fathollah Pour and Ashtiani (2022) developed a framework to study the generalization of classes of noisy hypotheses and show that adding noise to the output of neurons in a network can be helpful in generalization. Jim et al. (1996) show that additive and multiplicative noise can help speed up the convergence of RNNs on local minima surfaces. Recently, Lim et al. (2021) showed that noisy RNNs are more stable and robust to input perturbations by formalizing the regularization effects of noise.

Another line of work focuses on the generalization of neural network that are trained with Stochastic Gradient Descent (SGD) or its noisy variant Stochastic Gradient Langevin Descent (SGLD) (Russo and Zou, 2016; Xu and Raginsky, 2017; Russo and Zou, 2019; Steinke and Zakynthinou, 2020; Raginsky et al., 2017; Haghifam et al., 2020; Neu et al., 2021). Zhao et al. (2020) analyze the memory properties of recurrent networks and how well they can remember the input sequence.Tu et al. (2020) study the generalization of RNN by analyzing the Fisher-Rao norm of weights, which they obtain from the gradients of the network. They offer generalization bounds that can potentially become polynomial in T. Allen-Zhu and Li (2019) analyze the change in output through the dynamics of training RNNs and prove generalization bounds for recurrent networks that are again polynomial in T.

B Miscellaneous facts

Lemma 27 (Data Processing Inequality for TV Distance)**.** *Given two random variables $\overline{x_1}, \overline{x_2} \in \overline{\mathcal{X}}$, and a (random) Borel function $f : \mathcal{X} \to \mathcal{Y}$,*

$$d_{TV}(f(\overline{x_1}), f(\overline{x_2})) \leq d_{TV}(\overline{x_1}, \overline{x_2}).$$

Lemma 28. *Let $\overline{x}, \overline{y} \in \overline{\mathcal{X}}$ be two random variables with probability measures μ and ν. Denote by $\Pi(\mu, \nu)$ the set of all their couplings. Then, there exists $\pi^* \in \Pi(\mu, \nu)$ such that $\mathbb{P}_{\pi^*}[\overline{x} \neq \overline{y}] = d_{TV}(\mu, \nu)$, where the subscript π^* signals that the probability law is associated with the coupling π^*. Moreover, for any coupling $\pi \in \Pi(\mu, \nu)$ we have $\mathbb{P}_{\pi}[\overline{x} \neq \overline{y}] \geq d_{TV}(\mu, \nu)$.*

We use the following two lemmas to reason about the covering number of our recurrent model when we take the first dimensions of the output at each time t and when we concatenate new inputs with the outputs. The first lemma states that if two random variables are close to each other with respect to total variation distance, then they are still close after the applications of the $\text{First}(.)$ and $\text{Last}(.)$ functions.

Lemma 29 (From TV of Random Variable to TV of First and Last)**.** *Let $\overline{x_1}, \overline{x_2} \in \mathbb{R}^d$ be two random variables. We have*

$$d_{TV}\left(\textit{First}\left(\overline{x_1}\right), \textit{First}\left(\overline{x_2}\right)\right) \leq d_{TV}\left(\overline{x_1}, \overline{x_2}\right),$$
$$d_{TV}\left(\textit{Last}\left(\overline{x_1}\right), \textit{Last}\left(\overline{x_2}\right)\right) \leq d_{TV}\left(\overline{x_1}, \overline{x_2}\right).$$

Proof. We know that $\text{First}(.)$ and $\text{Last}(.)$ are functions from $\mathbb{R}^d$ to $\mathbb{R}^{d-1}$. Therefore we can apply Lemma 27 and conclude the result. □

The next lemma is used to bound the total variation distance between two random variables after being concatenated with the input at time t. In that case, we let $\overline{x_1}$ and $\overline{x_2}$ in the lemma to be $\text{First}\left(f^R\left(U, t-1\right)\right)$ and $\text{First}\left(\hat{f}^R\left(U, t-1\right)\right)$, which are in $\overline{\mathcal{X}_{p_k-1}}$. We also let $\overline{y}$ be $u^{(t)} \in \overline{\Delta_d}$, which is the input at time t.

Lemma 30 (From TV of Random Variable to TV of Concatenation)**.** *Let $\overline{x_1}, \overline{x_2}$ be random variables in $\overline{\mathcal{X}_d}$. Further, let $\overline{y}$ a random variable in $\overline{\Delta_d}$. If we have $d_{TV}\left(\overline{x_1}, \overline{x_2}\right) \leq \epsilon$, then*

$$d_{TV}\left(\left[\overline{x_1} \quad \overline{y}\right]^\top, \left[\overline{x_2} \quad \overline{y}\right]^\top\right) \leq \epsilon.$$

Proof. Let $y \in \overline{\Delta_d}$ be the random variable with Dirac delta measure on y_0. From Lemma 28 we know that there exists a maximal coupling π^* of $\overline{x_1}$ and $\overline{x_2}$ such that $d_{TV}(\overline{x_1}, \overline{x_2}) = \mathbb{P}_{\pi^*}[\overline{x_1} \neq \overline{x_2}]$ and denote the density associated with $\mathbb{P}_{\pi^*}$ by f^*. Let γ be a coupling of $\left[\overline{x_1} \quad \overline{y_1}\right]^\top$ and $\left[\overline{x_2} \quad \overline{y_2}\right]^\top$ such that

$$\hat{f}\left(\left[x_1 \quad y_1\right]^\top, \left[x_2 \quad y_2\right]^\top\right) = \begin{cases} f^*\left(x_1, x_2\right) & y_1 = y_2 = y_0, \\ 0 & \text{otherwise.} \end{cases}$$

We can easily verify that γ is a valid coupling. Denote by $f_{x_1 y}$ the density of the random variable $\left[\overline{x_1} \quad y\right]^\top$. We know that

$$f_{\overline{x_1 y}}\left(\left[x_1 \quad y_1\right]^\top\right) = \begin{cases} f_{\overline{x_1}}(x_1) & y = y_0, \\ 0 & \text{otherwise,} \end{cases}$$

where $f_{\overline{x_1}}$ is the density function of the random variable $\overline{x_1}$. We can observe that density associated with the marginal of γ would be the same as the density of the marginal of π^* at points where $y = y_0$ and it is zero otherwise. On the other hand, we know that π^* is a valid coupling of $\overline{x_1}$ and $\overline{x_2}$ and therefore the density of its marginal is $f_{\overline{x_1}}$. This concludes that the density of the marginal of γ is indeed $f_{\overline{x_1 y}}$. We can show the similar thing for the other marginal, which concludes that γ is a valid coupling.

Therefore, from Lemma 28 we can write that

$$
\begin{aligned}
& d_{TV}\left(\left[\overline{x_1} \quad \overline{y}\right]^\top, \left[\overline{x_2} \quad \overline{y}\right]^\top\right) \leq \mathbb{P}_\gamma\left[\left[\overline{x_1} \quad \overline{y}\right]^\top \neq \left[\overline{x_2} \quad \overline{y}\right]^\top\right] \\
& \leq \int_{\begin{bmatrix} x_1 \\ y \end{bmatrix} \neq \begin{bmatrix} x_2 \\ y \end{bmatrix}} \hat{f}\left(\left[x_1 \quad y\right]^\top, \left[x_2 \quad y\right]^\top\right) \leq \int_{\substack{x_1 \neq x_2, \\ y = y_0}} \hat{f}\left(\left[x_1 \quad y\right]^\top, \left[x_2 \quad y\right]^\top\right) \\
& \leq \int_{\substack{x_1 \neq x_2, \\ y = y_0}} f^*\left(x_1, x_2\right) \leq \mathbb{P}_{\pi^*}\left[\overline{x_1} \neq \overline{x_2}\right] = d_{TV}\left(\overline{x_1}, \overline{x_2}\right) \leq \epsilon.
\end{aligned}
$$

□

C Proof of lower bound

In order to prove Theorem 10, we first need to give the definition of PAC Learning with respect to $0-1$ loss.

Definition 31 (Agnostic PAC Learning with Respect to 0-1 Loss)**.** *We say that a hypothesis class $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ is agnostic PAC learnable with respect to $0-1$ loss if there exists a learner $\mathcal{A}$ and a function $m^{0-1}:(0,1)^2 \to \mathbb{N}$ with the following property: For every distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1,1\}$ and every $\epsilon, \delta \in (0,1)$, if S is a set of $m(\epsilon,\delta)$ i.i.d. samples from $\mathcal{D}$, then with probability at least $1-\delta$ (over the randomness of S) we have*

$$\mathbb{E}_{(x,y)\sim D}\left[l^{0-1}(\mathcal{A}(S),x,y)\right] \leq \inf_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim D}\left[l^{0-1}(f,x,y)\right] + \epsilon.$$

Same as Definition 9, we denote by $m_{\mathcal{F}}^{0-1}(\epsilon,\delta)$ the *sample complexity* of PAC learning $\mathcal{F}$ with respect to $0-1$ loss, which is the minimum number of samples required for learning $\mathcal{F}$ among all learners $\mathcal{A}$.

Before proving Theorem 10, we first prove Lemma 11, which, as mentioned before, is a core part of the proof. We state the lemma once more for completeness.

Lemma 32. *Let $\mathcal{H}_w$ = REC[MNET[p_0, p_k, w], T] be a well-defined class and let $\mathcal{F}_w = \{f : [-1/2,1/2]^{p\times T} \to \{-1,1\} \mid f(U) = sign\,(h(U)), h \in \mathcal{H}_w\}$. Then, for every distribution $\mathcal{D}$ over $[-1/2,1/2]^{p\times T} \times \{-1,1\}$, $\eta > 0$, and every function $f \in \mathcal{F}_w$ there exists a function $h \in \mathcal{H}_w$ such that $\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma\left(h,U,y\right)\right] \leq \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l^{0-1}\left(f,U,y\right)\right] + \eta$ where $l^{0-1}(f,U,y) = 1\left\{f(U) \neq y\right\}$.*

Proof. We know that $\mathcal{H}_w = \{h : \mathbb{R}^{p\times T} \to [-1/2,1/2] \mid h(u) = \text{Last}\left(b^R\left(U,T-1\right)\right), b \in \text{MNET}[p_0,p_k,w]\}$. Similarly, $\mathcal{F}_w = \{f : \mathbb{R}^{p\times T} \to \{-1,1\} \mid f(u) = \text{sign}\left(\text{Last}\left(b^R\left(U,T-1\right)\right)\right), b \in \text{MNET}[p_0,p_k,w]\}$. Fix a distribution $\mathcal{D}$ over $[-1/2,1/2]^{p\times T} \times \{-1,1\}$. Define

$$z = \min_b \operatorname*{arg\,max}_{0<x<\frac{1}{2}} \mathbb{P}\left[\left|\text{Last}\left(b^R(U,T-1)\right)\right| \geq x\right] \geq 1-\eta,$$

where the minimum is taken over all well-defined multi-layer neural networks b in MNET[p_0,p_k,w]. The last dimension of function b is in $[-1/2,1/2]$ and, intuitively, z is the largest possible value such that $\mathbb{P}\left[-z < \text{Last}\left(b^R(U,T-1)\right) < z\right] < \eta$.

Let f be any function in $\mathcal{F}_w$ and let $b = b_{k-1} \circ \ldots \circ b_0$ be the k-layer network associated with f where b_i's are single-layer sigmoid neural networks, i.e., $f(U) = \text{sign}\left(\text{Last}\left(b^R\left(U,T-1\right)\right)\right)$. Let $W_{k-1} = \left[v_1 \quad \ldots \quad v_{p_k}\right]^\top$ be the weight matrix associated with b_{k-1}. Denote by $\hat{W}_{k-1} = \left[v_1 \quad \ldots \quad c.v_{p_k}\right]^\top$ the matrix that is exactly the same as W_{k-1} but every element in its last row is multiplied by $c = \phi^{-1}(\gamma)/\phi^{-1}(z)$. Note that $z > 0$ and, therefore, $\phi^{-1}(z) > 0$. Let $\hat{b}_{k-1}$ be the single-layer neural network that is defined by weight matrix $\hat{W}_{k-1}$, i.e., $\hat{b}_{k-1}(x) = \Phi\left(\hat{W}_{k-1}^\top x\right)$. Denote $\hat{b} = \hat{b}_{k-1} \circ \ldots \circ b_0$ and let $h(U) = \text{Last}\left(\hat{b}^R\left(U,T-1\right)\right)$ for any $U \in \mathbb{R}^{p\times T}$. Clearly, $\hat{b} \in \text{MNET}[p_0,p_k,w]$

and $h \in \mathcal{H}_w$. We claim that $\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma\left(h,U,y\right)\right] \leq \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l^{0-1}\left(f,U,y\right)\right] + \eta$. We can write the definition of ramp loss as

$$\begin{aligned}
&\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma\left(h,U,y\right)\right] = \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[r_\gamma\left(-h(U).y\right)\right] \\
&= \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[r_\gamma\left(-h(U).y\right)\middle|\,|h(U)| \geq \phi\left(c.\phi^{-1}\left(z\right)\right)\right].\mathbb{P}\left[|h(U)| \geq \phi\left(c.\phi^{-1}\left(z\right)\right)\right] \\
&+ \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[r_\gamma\left(-h(U).y\right)\middle|\,|h(U)| < \phi\left(c.\phi^{-1}\left(z\right)\right)\right].\mathbb{P}\left[|h(U)| < \phi\left(c.\phi^{-1}\left(z\right)\right)\right] \\
&= \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[r_\gamma\left(-h(U).y\right)\middle|\,|h(U)| \geq \gamma\right].\mathbb{P}\left[|h(U)| \geq \gamma\right] \\
&+ \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[r_\gamma\left(-h(U).y\right)\middle|\,|h(U)| < \gamma\right].\mathbb{P}\left[|h(U)| < \gamma\right],
\end{aligned} \tag{1}$$

where we used the fact that sigmoid is a monotonic increasing function with a unique inverse and that $\phi\left(c.\phi^{-1}(z)\right) = \phi\left(\phi^{-1}(\gamma)\right) = \gamma$. Notice that whenever $|h(U)| \geq \gamma$ we can also conclude that either $h(U).y \geq \gamma$ or $h(U).y \leq -\gamma$. This means that $r_\gamma\left(-h(U).y\right)$ is either 0 or 1. When $h(U).y \geq \gamma$ we have $r_\gamma\left(-h(U).y\right) = 0$ and when $h(U).y \leq -\gamma$ we have $r_\gamma\left(-h(U).y\right) = 1$. In other words if $|h(U)| \geq \gamma$, we have

$$r_\gamma\left(-h(U).y\right) = 1\left\{\text{sign}\left(h(U)\right) \neq y\right\} \tag{2}$$

On the other hand, we know that $\gamma, z > 0$ and $c = \phi^{-1}(\gamma)/\phi^{-1}(z) > 0$. Consequently, $\text{sign}\left(h(U)\right) = \text{sign}\left(\text{Last}\left(\hat{b}^R\left(U, T-1\right)\right)\right) = f(U)$ for any $U \in \mathbb{R}^{p\times T}$. Lemma 36 suggests that

$$\mathbb{P}\left[\left|\text{Last}\left(b^R(U,T-1)\right)\right| < z\right] = \mathbb{P}\left[\left|\text{Last}\left(\hat{b}^R(U,T-1)\right)\right| < \phi\left(c.\phi^{-1}(z)\right)\right] = \mathbb{P}\left[|h(U)| < \gamma\right].$$

Moreover, we know that z is chosen such that $\mathbb{P}\left[\left|\text{Last}\left(b^R(U,T-1)\right)\right| < z\right] < \eta$ and the ramp loss is at most 1. Taking this and Equations 1 and 2 into account we can write that

$$\begin{aligned}
&\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma\left(h,U,y\right)\right] = \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[1\left\{\text{sign}\left(h(U)\right) \neq y\right\}\middle|\,|h(U)| \geq \gamma\right].\mathbb{P}\left[|h(U)| \geq \gamma\right] \\
&+ \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[r_\gamma\left(-h(U).y\right)\middle||h(U)| < \gamma\right].\mathbb{P}\left[\left|\text{Last}\left(b^R(U,T-1)\right)\right| < z\right] \\
&\leq \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[1\left\{\text{sign}\left(h(U)\right) \neq y\right\}\middle|\,|h(U)| \geq \gamma\right].\mathbb{P}\left[|h(U)| \geq \gamma\right] + \eta \\
&\leq \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[1\left\{\text{sign}\left(h(U)\right) \neq y\right\}\middle|\,|h(U)| \geq \gamma\right].\mathbb{P}\left[|h(U)| \geq \gamma\right] \\
&+ \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[1\left\{\text{sign}\left(h(U)\right) \neq y\right\}\middle|\,|h(U)| < \gamma\right].\mathbb{P}\left[|h(U)| < \gamma\right] + \eta \\
&\leq \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[1\left\{\text{sign}\left(h(U)\right) \neq y\right\}\right] + \eta \\
&\leq \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l^{0-1}\left(f,U,y\right)\right] + \eta.
\end{aligned}$$

□

Proof of Theorem 10.

Proof. Define $\mathcal{F}_w = \{f : [-1/2, 1/2] \to \{-1, 1\} \mid f(U) = \text{sign}\left(h(U)\right), h \in \mathcal{H}_w\}$ as the class of all sigmoid recurrent networks with w weights that output binary values. Let $\mathcal{D}$ be a distribution over $[-1/2, 1/2]^{p\times T} \times \{-1, 1\}$. From Lemma 11 we know that for every $f \in \mathcal{F}_w$ there exists a function $h \in \mathcal{H}_w$ such that $\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma\left(h,U,y\right)\right] \leq \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l^{0-1}\left(f,U,y\right)\right] + \eta$, where $\eta > 0$ is any small value. Therefore, we can write that

$$\inf_{h\in\mathcal{H}_w} \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma\left(h,U,y\right)\right] \leq \inf_{f\in\mathcal{F}_w} \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l^{0-1}\left(f,U,y\right)\right] + \eta. \tag{3}$$

Let $m_{\mathcal{H}_w}(\epsilon, \delta)$ denote the sample complexity of PAC learning $\mathcal{H}_w$ with respect to ramp loss. Therefore, there exists an algorithm $\mathcal{A}$ that receives a set S of $m \geq m_{\mathcal{H}_w}(\epsilon, \delta)$ i.i.d. samples from $\mathcal{D}$ and returns $\hat{h} = \mathcal{A}(S)$ such that with probability at least $1 - \delta$ we have

$$\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l^\gamma\left(\hat{h}, U, y\right)\right] \leq \inf_{h\in\mathcal{H}_w} \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l^\gamma\left(h,U,y\right)\right] + \epsilon.$$

Let $\hat{f}$ be a function in $\mathcal{F}_w$ such that $\hat{f}(U) = \text{sign}\left(\hat{h}(U)\right)$ for every $U \in [-1/2, 1/2]^{p\times T}$. Given the definitions of $0-1$ loss and ramp loss, it is easy to verify that $\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l\right]^{0-1}\left(\hat{f}, U, y\right) \leq$

$\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l\right]^{\gamma}(\hat{h},U,y)$. Taking this and Equation 3 into account, we can define a new algorithm $\mathcal{A}'$ that, given the set S, returns $\hat{f}\in\mathcal{F}_w$ such that with probability at least $1-\delta$ we have

$$\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l\right]^{0-1}(\hat{f},U,y)\leq\inf_{h\in\mathcal{H}_w}\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l\right]^{\gamma}(h,U,y)+\epsilon\leq\inf_{f\in\mathcal{F}_w}\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l\right]^{0-1}(f,U,y)+\epsilon+\eta.$$

This means that we have

$$m^{0-1}_{\mathcal{F}_w}(\epsilon+\eta,\delta)\leq m_{\mathcal{H}_w}(\epsilon,\delta). \tag{4}$$

On the other hand, from Theorem 34 we now that the VC-dimension of $\mathcal{F}_w$ is $\Omega(wT)$. Moreover, Theorem 33 suggests that

$$m^{0-1}_{\mathcal{F}_w}(\epsilon,\delta)=\Omega\left(\frac{wT+\log(1/\delta)}{\epsilon^2}\right).$$

Taking the above equation and Equation 4 into account, by setting $\eta=O(\epsilon)$, we can write that

$$m_{\mathcal{H}_w}(\epsilon,\delta)=\Omega\left(\frac{wT+\log(1/\delta)}{\epsilon^2}\right),$$

which concludes our result. □

The following theorem states that we can find a lower bound on the sample complexity of PAC learning $\mathcal{F}$ with respect to $0-1$ loss based on its VC-dimension. For a proof see Theorems 5.2, and 5.10 in Anthony et al. (1999).

Theorem 33 (Lower Bound on the Sample Complexity of PAC Learning (Anthony et al., 1999))**.** *Let $\mathcal{F}$ be a class of functions from a domain $\mathcal{X}$ to $\{-1,1\}$ and let $d=VC(\mathcal{F})$ be the VC-dimension of the class $\mathcal{F}$. Assume $d<\infty$. Then there exists an absolute constant C such that for every $(\epsilon,\delta)\in(0,1/40)$ we have*

$$m^{0-1}_{\mathcal{F}}(\epsilon,\delta)\geq C\frac{d+\log(1/\delta)}{\epsilon^2}.$$

We now introduce a lower bound on the VC-dimension of sigmoid recurrent neural networks with binary outputs which is based on a result due to Koiran and Sontag (1998).

Theorem 34 (A Lower Bound on VC-Dimension of Sigmoid Recurrent Neural Networks)**.** *For every $T\geq 3$ and $w\geq 19$ there exists a well-defined class $\mathcal{H}_w=$ REC[MNET[p_0,p_k,w], T] with the following property: The VC-dimension of $\mathcal{F}_w=\{f:[-1/2,1/2]^{p\times T}\to\{-1,1\}\mid f(U)=\text{sign}\left(h(U)\right),h\in\mathcal{H}_w\}$ is $\Omega\left(wT\right)$.*

The proof of the above theorem is essentially the same as the proof of the result in Koiran and Sontag (1998). The only difference is that the we should construct our network in a way that the last two dimensions of the output of MNET[p_0,p_k,w] must be similar to each other in order to feed back the value of Last $\left(b^R(f,t-1)\right)$ with an extra node. Therefore, we only need a network that has a constant factor more weights than the network that is proposed in Koiran and Sontag (1998) which does not change the order of sample complexity.

C.1 Lemmas used in the proof of Lemma 11

In the following, we state two lemmas that will help in proving Lemma 11.

Lemma 35. *Let $W_{k-1}=[v_1\dots v_{p_k}]\in\mathbb{R}^{p_{k-1}\times p_k}$ and $\hat{W}_{k-1}=\left[v_1\quad\dots\quad c.v_{p_k}\right]^{\top}$ for a constant $c>0$. Define two single-layer networks $b_{k-1}(x)=\Phi\left(W_{k-1}^{\top}x\right)$ and $\hat{b}_{k-1}(x)=\Phi\left(\hat{W}_{k-1}^{\top}x\right)$. Then, for any two multi-layer networks $b=b_{k-1}\circ\dots\circ b_0$ and $\hat{b}=\hat{b}_{k-1}\circ\dots\circ b_0$ in a well-defined class MNET[p_0,p_k,w], every $U\in[-1/2,1/2]^{p\times T}$, and every $t\in[T-1]$ we have*

$$\textit{First}\left(b^R\left(U,t\right)\right)=\textit{First}\left(\hat{b}^R\left(U,t\right)\right).$$

Proof. We prove by induction. Denote $r = b_{k-2} \circ \ldots \circ b_o$. Therefore, we have $b = b_{k-1} \circ r$ and $\hat{b} = \hat{b}_{k-1} \circ r$. For $t = 0$, we can denote $x^{(0)} = r\left(\begin{bmatrix} 0_{q-1} & u^{(0)} \end{bmatrix}^\top\right)$ and write that

$$\begin{aligned} &\text{First}\left(b^R(U,0)\right) = \text{First}\left(b_{k-1}\left(r\left(\begin{bmatrix} 0_{q-1} \\ u^{(0)} \end{bmatrix}\right)\right)\right) = \text{First}\left(b_{k-1}\left(x^{(0)}\right)\right) \\ &= \text{First}\left(\begin{bmatrix} \phi\left(\langle v_1, x^{(0)}\rangle\right) \\ \vdots \\ \phi\left(\langle v_{p_k-1}, x^{(0)}\rangle\right) \end{bmatrix}\right) = \text{First}\left(\begin{bmatrix} \phi\left(\langle v_1, x^{(0)}\rangle\right) \\ \vdots \\ \phi\left(\langle c.v_{p_k-1}, x^{(0)}\rangle\right) \end{bmatrix}\right) \\ &= \text{First}\left(\hat{b}_{k-1}\left(x^{(0)}\right)\right) = \text{First}\left(\hat{b}^R(U,0)\right), \end{aligned}$$

where $\langle v_i, x^{(t)}\rangle$ denotes the inner product between vectors v_i and $x^{(t)}$. Assume that we have $\text{First}\left(b^R(U,t-1)\right) = \text{First}\left(\hat{b}^R(U,t-1)\right)$ for $t-1 \in [T-2]$. We now prove that we also have $\text{First}\left(b^R(U,t)\right) = \text{First}\left(\hat{b}^R(U,t)\right)$. Denote $x^{(t)} = r\left(\begin{bmatrix} \text{First}\left(b^R(U,t-1)\right) & u^{(t)} \end{bmatrix}^\top\right)$. We can then write that

$$\begin{aligned} &\text{First}\left(b^R(U,t)\right) = \text{First}\left(b_{k-1} \circ r\left(\begin{bmatrix} \text{First}\left(b^R(U,t-1)\right) \\ u^{(t)} \end{bmatrix}\right)\right) = \text{First}\left(b_{k-1}\left(x^{(t)}\right)\right) \\ &= \text{First}\left(\begin{bmatrix} \phi\left(\langle v_1, x^{(t)}\rangle\right) \\ \vdots \\ \phi\left(\langle v_{p_k-1}, x^{(t)}\rangle\right) \end{bmatrix}\right) = \text{First}\left(\begin{bmatrix} \phi\left(\langle v_1, x^{(t)}\rangle\right) \\ \vdots \\ \phi\left(\langle c.v_{p_k-1}, x^{(t)}\rangle\right) \end{bmatrix}\right) \\ &= \text{First}\left(\hat{b}_{k-1}\left(x^{(t)}\right)\right) = \text{First}\left(\hat{b}^R(U,t)\right). \end{aligned}$$

□

Lemma 36. *Let* $W_{k-1} = [v_1 \ldots v_{p_k}] \in \mathbb{R}^{p_{k-1} \times p_k}$ *and* $\hat{W}_{k-1} = \begin{bmatrix} v_1 & \ldots & c.v_{p_k} \end{bmatrix}^\top$ *for a constant* $c > 0$. *Define two single-layer networks* $b_{k-1}(x) = \Phi\left(W_{k-1}^\top x\right)$ *and* $\hat{b}_{k-1}(x) = \Phi\left(\hat{W}_{k-1}^\top x\right)$. *Let* $\mathcal{D}$ *be a distribution over* $[-1/2, 1/2]^{p \times T}$. *Then, for any two multi-layer networks* $b = b_{k-1} \circ \ldots \circ b_0$ *and* $\hat{b} = \hat{b}_{k-1} \circ \ldots \circ b_0$ *in a well-defined class MNET[*p_0, p_k, w*] we have*

$$\mathbb{P}\left[\left|\textit{Last}\left(b^R(U,T-1)\right)\right| < z\right] = \mathbb{P}\left[\left|\textit{Last}\left(\hat{b}^R(U,T-1)\right)\right| < \phi\left(c.\phi^{-1}(z)\right)\right],$$

where $\phi^{-1}(z)$ *is the inverse of sigmoid function* ϕ *at* z.

Proof. Denote $r = b_{k-2} \circ \ldots \circ b_o$ and $x^{(T-1)} = r\left(\begin{bmatrix} \text{First}\left(b^R(U,T-2)\right) & u^{(T-1)} \end{bmatrix}^\top\right)$. Note that

$$\begin{aligned} &\text{Last}\left(b^R(U,T-1)\right) = \text{Last}\left(b_{k-1} \circ r\left(\begin{bmatrix} \text{First}\left(b^R(U,T-2)\right) \\ u^{(T-1)} \end{bmatrix}\right)\right) \\ &= \text{Last}\left(b_{k-1}\left(x^{(T-1)}\right)\right) = \phi\left(\langle v_{p_k}, x^{(T-1)}\rangle\right), \end{aligned}$$

where $\langle v_{p_k}, x^{(T-1)}\rangle$ denotes the inner product between v_{p_k} and $x^{(T-1)}$. From Lemma 35, we know that $\text{First}\left(b^R(U,T-2)\right) = \text{First}\left(\hat{b}^R(U,T-2)\right)$. Therefore, we also have that

$$\begin{aligned} &\text{Last}\left(\hat{b}^R(U,T-1)\right) = \text{Last}\left(\hat{b}_{k-1} \circ r\left(\begin{bmatrix} \text{First}\left(\hat{b}^R(U,T-2)\right) \\ u^{(T-1)} \end{bmatrix}\right)\right) \\ &= \text{Last}\left(\hat{b}_{k-1}\left(x^{(T-1)}\right)\right) = \phi\left(\langle c.v_{p_k}, x^{(T-1)}\rangle\right). \end{aligned}$$

Considering the above equations and the facts that $\phi(x)$ is an invertible and strictly increasing function and that $\phi(x) = -\phi(-x)$, we can write

$$\begin{aligned}
&\mathbb{P}\left[\left|\text{Last}\left(b^R\left(U,T-1\right)\right)\right| < z\right] = \mathbb{P}\left[-z < \text{Last}\left(b^R\left(U,T-1\right)\right) < z\right] \\
&= \mathbb{P}\left[-z \leq \phi\left(\langle v_{p_k}, x^{(T-1)}\rangle\right) < z\right] = \mathbb{P}\left[\phi^{-1}(-z) < \langle v_{p_k}, x^{(T-1)}\rangle < \phi^{-1}\left(z\right)\right] \\
&= \mathbb{P}\left[-c.\phi^{-1}(z) \leq \langle c.v_{p_k}, x^{(T-1)}\rangle < c.\phi^{-1}\left(z\right)\right] \\
&= \mathbb{P}\left[\phi\left(-c.\phi^{-1}\left(z\right)\right) < \phi\left(\langle c.v_{p_k}, x^{(T-1)}\rangle\right) < \phi\left(c.\phi^{-1}\left(z\right)\right)\right] \\
&= \mathbb{P}\left[-\phi\left(c.\phi^{-1}\left(z\right)\right) < \phi\left(\langle c.v_{p_k}, x^{(T-1)}\rangle\right) < \phi\left(c.\phi^{-1}\left(z\right)\right)\right] \\
&= \mathbb{P}\left[\left|\text{Last}\left(\hat{b}^R\left(U,T-1\right)\right)\right| < \phi\left(c.\phi^{-1}\left(z\right)\right)\right].
\end{aligned}$$

□

D Proof of upper bound

D.1 Proof of Theorem 24

We prove the following general theorem which holds for input domains $\overline{\mathcal{X}_s}$ and $\Delta_{p\times T}$.

Theorem 37 (TV Covering Number of $\overline{\mathcal{G}_\sigma} \circ \text{REC}[\overline{\mathcal{F}_\sigma}, T]$ From $\overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{F}_\sigma}$)**.** *Let $s, p, q \in \mathbb{N}$ such that $s = p + q - 1$. Let $\overline{\mathcal{F}}$ be a class of functions from $\overline{\mathcal{X}_s}$ to $\overline{\mathcal{X}_q}$ and denote by $\overline{\mathcal{F}_\sigma} = \overline{\mathcal{F}} \circ \overline{\mathcal{G}_{\sigma,s}}$ the class of its composition with noise. Then we have*

$$N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ \mathit{REC}[\overline{\mathcal{F}_\sigma}, T], \infty, d_{TV}^\infty, \overline{\Delta_{p\times T}}\right) \leq N_U\left(\epsilon/T, \overline{\mathcal{G}_{\sigma,q}} \circ \overline{\mathcal{F}_\sigma}, \infty, d_{TV}^\infty, \overline{\mathcal{X}_s}\right).$$

Proof. Let $C = \{\overline{g_{\sigma,q}} \circ \hat{f}_i \circ \overline{g_{\sigma,s}} \mid \hat{f}_i \circ \overline{g_{\sigma,s}} \in \overline{\mathcal{F}_\sigma}, i \in [r]\}$ be a global ϵ-cover for $\overline{\mathcal{G}_{\sigma,s}} \circ \overline{\mathcal{F}_\sigma}$ with respect to domain $\overline{\mathcal{X}_s}$ and d_{TV}^∞. Therefore, $|C| \leq N_U\left(\epsilon, \overline{\mathcal{G}_{\sigma,q}} \circ \overline{\mathcal{F}_\sigma}, \infty, d_{TV}^\infty, \overline{\mathcal{X}_s}\right)$. Then for any function $\overline{g_{\sigma,q}} \circ f \circ \overline{g_{\sigma,s}} \in \overline{\mathcal{G}_{\sigma,q}} \circ \overline{\mathcal{F}_\sigma}$ we know that there exists a function $\overline{g_{\sigma,q}} \circ \hat{f}_i \circ \overline{g_{\sigma,s}}$ in C such that for every $\overline{x} \in \overline{\mathcal{X}_s}$ we have $d_{TV}\left(\overline{g_{\sigma,q}} \circ f \circ \overline{g_{\sigma,s}}(\overline{x}), \overline{g_{\sigma,q}} \circ \hat{f}_i \circ \overline{g_{\sigma,s}}(\overline{x})\right) \leq \epsilon$. Denote $\overline{h} = f \circ \overline{g_{\sigma,s}}$ and $\overline{\hat{h}_i} = \hat{f}_i \circ \overline{g_{\sigma,s}}$. We prove by induction that for any input matrix $\overline{U} = \left[\overline{u^{(0)}} \quad \ldots \quad \overline{u^{(T-1)}}\right] \in \overline{\Delta_{p\times T}}$, where $\overline{u^{(t)}} = \overline{\delta_{u^{(t)}}}$, we have $d_{TV}\left(\overline{g_{\sigma,q}} \circ \overline{h}^R\left(\overline{U}, T-1\right), \overline{g_{\sigma,q}} \circ \overline{\hat{h}_i}^R\left(\overline{U}, T-1\right)\right) \leq T\epsilon$.

We start by proving that $d_{TV}\left(\overline{g_{\sigma,q}} \circ \overline{h}^R\left(\overline{U}, 0\right), \overline{g_{\sigma,q}} \circ \overline{\hat{h}_i}^R\left(\overline{U}, 0\right)\right) \leq \epsilon$. Denote $\overline{x^{(0)}} = \left[\overline{\delta_{0_{q-1}}} \quad \overline{u^{(0)}}\right]^\top \in \overline{\Delta_s}$. We can write that

$$\begin{aligned}
&d_{TV}\left(\overline{g_{\sigma,q}} \circ h^R\left(\overline{U}, 0\right), \overline{g_{\sigma,q}} \circ \hat{h}_i^R\left(\overline{U}, 0\right)\right) \\
&= d_{TV}\left(\overline{g_{\sigma,q}} \circ f \circ \overline{g_{\sigma,s}}\left(\begin{bmatrix}\overline{\delta_{0_{q-1}}} \\ \overline{u^{(0)}}\end{bmatrix}\right), \overline{g_{\sigma,q}} \circ \hat{f}_i \circ \overline{g_{\sigma,s}}\left(\begin{bmatrix}\overline{\delta_{0_{q-1}}} \\ \overline{u^{(0)}}\end{bmatrix}\right)\right).
\end{aligned}$$

Since $\left(\left[\overline{\delta_{0_{q-1}}} \quad \overline{u^{(0)}}\right]^\top\right) \in \overline{\mathcal{X}_s}$ and considering the fact that $\overline{g_{\sigma,q}} \circ f \circ \overline{g_{\sigma,s}} = \overline{g_{\sigma,q}} \circ h \in \overline{\mathcal{G}_{\sigma,q}} \circ \overline{\mathcal{F}_\sigma}$ and $\overline{g_{\sigma,q}} \circ \hat{f}_i \circ \overline{g_{\sigma,s}} = \overline{g_{\sigma,q}} \circ \overline{\hat{h}_i} \in \overline{\mathcal{G}_{\sigma,q}} \circ \overline{\mathcal{F}_\sigma}$ are globally ϵ-close over $\overline{\mathcal{X}_s}$, we get that

$$d_{TV}\left(\overline{g_{\sigma,q}} \circ \overline{h}^R\left(\overline{U}, 0\right), \overline{g_{\sigma,q}} \circ \overline{\hat{h}_i}^R\left(\overline{U}, 0\right)\right) \leq \epsilon.$$

Now assume that we have

$$d_{TV}\left(\overline{g_{\sigma,q}} \circ \overline{h}^R\left(\overline{U}, t-1\right), \overline{g_{\sigma,q}} \circ \overline{\hat{h}_i}^R\left(\overline{U}, t-1\right)\right) \leq t\epsilon. \tag{5}$$

We want to bound the total variation distance between $\overline{g_{\sigma,q}} \circ \overline{h}^R\left(\overline{U}, t\right)$ and $\overline{g_{\sigma,q}} \circ \overline{\hat{h}_i}^R\left(\overline{U}, t\right)$, which are defined as follows.

$$\overline{g_{\sigma,q}} \circ \overline{h}^R\left(\overline{U}, t\right) = \overline{g_{\sigma,q}} \circ f \circ \overline{g_{\sigma,s}}\left(\begin{bmatrix}\text{First}\left(\overline{h}^R\left(\overline{U}, t-1\right)\right) \\ \overline{u^{(t)}}\end{bmatrix}\right),$$
$$\overline{g_{\sigma,q}} \circ \overline{\hat{h}_i}^R\left(\overline{U}, t\right) = \overline{g_{\sigma,q}} \circ \hat{f}_i \circ \overline{g_{\sigma,s}}\left(\begin{bmatrix}\text{First}\left(\overline{\hat{h}_i}^R\left(\overline{U}, t-1\right)\right) \\ \overline{u^{(t)}}\end{bmatrix}\right).$$

From Lemma 29 we know that

$$\begin{aligned}&d_{TV}\left(\text{First}\left(\overline{g_{\sigma,q}}\left(\overline{h}^R\left(\overline{U}, t-1\right)\right)\right), \text{First}\left(\overline{g_{\sigma,q}}\left(\overline{\hat{h}_i}^R\left(\overline{U}, t-1\right)\right)\right)\right) \\ &\leq d_{TV}\left(\overline{g_{\sigma,q}}\left(\overline{h}^R\left(\overline{U}, t-1\right)\right), \overline{g_{\sigma,q}}\left(\overline{\hat{h}_i}^R\left(\overline{U}, t-1\right)\right)\right) \leq t\epsilon\end{aligned}$$

It is easy to verify that $\text{First}\left(\overline{g_{\sigma,q}}\left(\overline{h}^R\left(\overline{U}, t-1\right)\right)\right) = \overline{g_{\sigma,q-1}}\left(\text{First}\left(\overline{h}^R\left(\overline{U}, t-1\right)\right)\right)$ because $\overline{g_{\sigma,q}}$ is a gaussian noise with covariance matrix equal to $\sigma^2 I_q$, where $I_q \in \mathbb{R}^{q \times q}$ is the identity matrix. Considering this fact and Lemma 30 we can write that

$$\begin{aligned}&d_{TV}\left(\begin{bmatrix}\overline{g_{\sigma,q-1}}\left(\text{First}\left(\overline{h}^R\left(\overline{U}, t-1\right)\right)\right) \\ \overline{u^{(t)}}\end{bmatrix}, \begin{bmatrix}\overline{g_{\sigma,q-1}}\left(\text{First}\left(\overline{\hat{h}_i}^R\left(\overline{U}, t-1\right)\right)\right) \\ \overline{u^{(t)}}\end{bmatrix}\right) \\ \leq &d_{TV}\left(\text{First}\left(\overline{g_{\sigma,q}}\left(\overline{h}^R\left(\overline{U}, t-1\right)\right)\right), \text{First}\left(\overline{g_{\sigma,q}}\left(\overline{\hat{h}_i}^R\left(\overline{U}, t-1\right)\right)\right)\right) \leq t\epsilon.\end{aligned}$$

Applying data processing inequality for TV distance (i.e., Lemma 27) we can write that

$$\begin{aligned}&d_{TV}\left(\begin{bmatrix}\overline{g_{\sigma,q-1}}\left(\text{First}\left(\overline{h}^R\left(\overline{U}, t-1\right)\right)\right) \\ \overline{g_{\sigma,p}}\left(\overline{u^{(t)}}\right)\end{bmatrix}, \begin{bmatrix}\overline{g_{\sigma,q-1}}\left(\text{First}\left(\overline{\hat{h}_i}^R\left(\overline{U}, t-1\right)\right)\right) \\ \overline{g_{\sigma,p}}\left(\overline{u^{(t)}}\right)\end{bmatrix}\right) \\ &= d_{TV}\left(\overline{g_{\sigma,s}}\left(\begin{bmatrix}\text{First}\left(\overline{h}^R\left(\overline{U}, t-1\right)\right) \\ \overline{u^{(t)}}\end{bmatrix}\right), \overline{g_{\sigma,s}}\left(\begin{bmatrix}\text{First}\left(\overline{\hat{h}_i}^R\left(\overline{U}, t-1\right)\right) \\ \overline{u^{(t)}}\end{bmatrix}\right)\right) \leq t\epsilon.\end{aligned} \tag{6}$$

Notice that $\begin{bmatrix}\text{First}\left(\overline{h}^R\left(\overline{U}, t-1\right)\right) & \overline{u^{(t)}}\end{bmatrix}^\top$ is in $\overline{\mathcal{X}_s}$. Since we know that $\overline{g_{\sigma,q}} \circ f \circ \overline{g_{\sigma,s}}$ and $\overline{g_{\sigma,q}} \circ \hat{f}_i \circ \overline{g_{\sigma,s}}$ are globally ϵ-close on $\overline{\mathcal{X}_s}$, we can write that

$$d_{TV}\left(\overline{g_{\sigma,q}} \circ f \circ \overline{g_{\sigma,s}}\left(\begin{bmatrix}\text{First}\left(\overline{h}^R\left(\overline{U}, t-1\right)\right) \\ \overline{u^{(t)}}\end{bmatrix}\right), \overline{g_{\sigma,q}} \circ \hat{f}_i \circ \overline{g_{\sigma,s}}\left(\begin{bmatrix}\text{First}\left(\overline{h}^R\left(\overline{U}, t-1\right)\right) \\ \overline{u^{(t)}}\end{bmatrix}\right)\right) \leq \epsilon. \tag{7}$$

Moreover, from data processing inequality (i.e., Lemma 27) and Equation 6 we can conclude that

$$d_{TV}\left(\overline{g_{\sigma,q}} \circ \overline{\hat{f}_i} \circ \overline{g_{\sigma,s}}\left(\begin{bmatrix}\text{First}\left(\overline{h}^R\left(\overline{U}, t-1\right)\right) \\ \overline{u^{(t)}}\end{bmatrix}\right), \overline{g_{\sigma,q}} \circ \hat{f}_i \circ \overline{g_{\sigma,s}}\left(\begin{bmatrix}\text{First}\left(\overline{\hat{h}_i}^R\left(\overline{U}, t-1\right)\right) \\ \overline{u^{(t)}}\end{bmatrix}\right)\right) \leq t\epsilon. \tag{8}$$

Finally, we can combine Equations 7 and 8 together with the triangle inequality for total variation distance to conclude that

$$\begin{aligned}&d_{TV}\left(\overline{g_{\sigma,q}} \circ f \circ \overline{g_{\sigma,s}}\left(\begin{bmatrix}\text{First}\left(\overline{h}^R\left(\overline{U}, t-1\right)\right) \\ \overline{u^{(t)}}\end{bmatrix}\right), \overline{g_{\sigma,q}} \circ \hat{f}_i \circ \overline{g_{\sigma,s}}\left(\begin{bmatrix}\text{First}\left(\overline{\hat{h}_i}^R\left(\overline{U}, t-1\right)\right) \\ \overline{u^{(t)}}\end{bmatrix}\right)\right) \\ &= d_{TV}\left(\overline{g_{\sigma,q}} \circ \overline{h}^R\left(\overline{U}, t\right), \overline{g_{\sigma,q}} \circ \overline{\hat{h}_i}^R\left(\overline{U}, t\right)\right) \leq (t+1)\epsilon.\end{aligned}$$

So far, we have proved that for any input matrix $\overline{U} \in \overline{\Delta_{p\times T}}$ we have

$$d_{TV}\left(\overline{g_{\sigma,q}} \circ \overline{h}^{R}\left(\overline{U}, T-1\right), \overline{g_{\sigma,q}} \circ \overline{\hat{h}_i}^{R}\left(\overline{U}, T-1\right)\right) \leq T\epsilon$$

By another application of Lemma 29 we can conclude that

$$d_{TV}\left(\text{Last}\left(\overline{g_{\sigma,q}} \circ \overline{h}^{R}\left(\overline{U}, T-1\right)\right), \text{Last}\left(\overline{g_{\sigma,q}} \circ \overline{\hat{h}_i}^{R}\left(\overline{U}, T-1\right)\right)\right) \leq T\epsilon$$

We can have a similar argument to the first function and write the above equation as

$$d_{TV}\left(\overline{g_{\sigma,1}} \circ \text{Last}\left(h^{R}\left(\overline{U}, T-1\right)\right), \overline{g_{\sigma,1}} \circ \text{Last}\left(\hat{h}_i^{R}\left(\overline{U}, T-1\right)\right)\right) \leq T\epsilon.$$

This means that for every function $\overline{g_{\sigma,1}} \circ \text{Last}\left(\overline{h}^{R}\left(\overline{U}, T-1\right)\right)$ in $\overline{\mathcal{G}_{\sigma,1}} \circ \text{REC}[\overline{\mathcal{F}_\sigma}, T]$ there exists a function $\hat{f}_i$ in $\mathcal{F}$ such that $\overline{g_{\sigma,1}} \circ \text{Last}\left(\overline{h}^{R}\left(\overline{U}, T-1\right)\right)$ and $\overline{g_{\sigma,1}} \circ \text{Last}\left(\overline{\hat{h}_i}^{R}\left(\overline{U}, T-1\right)\right)$ are globally $T\epsilon$-cover close to each other with respect to $\overline{\Delta_{p\times T}}$. Setting $\epsilon' = \epsilon/T$ we can conclude that

$$N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ \text{REC}[\overline{\mathcal{F}_\sigma}, T], \infty, d_{TV}^{\infty}, \overline{\Delta_{p\times T}}\right) \leq N_U\left(\frac{\epsilon}{T}, \overline{\mathcal{G}_{\sigma,q}} \circ \overline{\mathcal{F}_\sigma}, \infty, d_{TV}^{\infty}, \overline{\mathcal{X}_s}\right).$$

The proof of the bounded domains essentially follows the same steps as above but for inputs that are bounded, i.e., inputs in $\overline{\Delta_{B,p\times T}}$ and $\overline{\mathcal{X}_{B,s}}$. □

D.2 A bound on the TV covering number of multi-layer noisy networks

From Theorem 25 and Theorem 23 we can get the following bound on the total variation covering number of noisy multi-layer networks.

Theorem 38 (TV Cover for Multi-Layer Noisy Neural Networks)**.** *For every $\epsilon, \sigma \in (0,1)$ and every well-defined class $\overline{MNET_\sigma}[p_0, p_k, w]$, we have*

$$\begin{aligned}&\log N_U(\epsilon, \overline{\mathcal{G}_{\sigma,p_k}} \circ \overline{MNET_\sigma}[p_0, p_k, w], \infty, d_{TV}^{\infty}, \overline{\mathcal{X}_{0.5,p_0}})\\ &= O\left(w \log\left(\frac{w}{\epsilon\sigma}\log\left(\frac{w}{\epsilon\sigma}\right)\right)\right) = \widetilde{O}\left(w\log\left(\frac{1}{\epsilon\sigma}\right)\right),\end{aligned}$$

where $\widetilde{O}$ hides logarithmic factors.

Proof. Fix a choice of $p_1, \dots, p_{k-1} \in \mathbb{N}$ and let $\overline{\mathcal{F}} = \text{NET}[p_{k-1}, p_k] \circ \dots \circ \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_0, p_1] \circ \overline{\mathcal{G}_\sigma}$ be a class of multi-layer sigmoid neural networks in $\overline{\text{MNET}_\sigma}[p_0, p_k, w]$. Notice that

$$\overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{F}} = \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_{k-1}, p_k] \circ \dots \circ \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_0, p_1] \circ \overline{\mathcal{G}_\sigma}$$

and that the covering number of $\overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{F}}$ with respect to $\overline{\mathcal{X}_{1,p_0}}$ is the same as the covering number of $\overline{\mathcal{G}_\sigma} \circ \text{NET}[p_{k-1}, p_k] \circ \dots \circ \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_0, p_1]$ with respect to $\overline{\mathcal{G}_{\sigma,p_0}} \circ \overline{\mathcal{X}_{1,p_0}}$. From Theorem 25 we know that for any $0 \leq i \leq k-1$ we can bound the covering number of $\overline{\mathcal{G}_\sigma} \circ \text{NET}[p_i, p_{i+1}]$ as

$$\begin{aligned}&\log N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_i, p_{i+1}], \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{0.5,p_i}}\right)\\ &\leq p_i(p_{i+1}+1)\log\left(30\frac{p_i^{5/2}\sqrt{\ln\left((5p_i - \epsilon\sigma)/(\epsilon\sigma)\right)}}{\epsilon^{3/2}\sigma^2}\ln\left(\frac{5p_i}{\epsilon\sigma}\right)\right).\end{aligned}$$

Note that in Fathollah Pour and Ashtiani (2022) the above bound was originally stated as a bound on the covering number of $\overline{\mathcal{G}_\sigma} \circ \text{NET}[p_i, p_{i+1}]$ with respect $\overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{1,p_i}}$. However, we know that the bound with respect to $\overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{1,p_i}}$ is always an upper bound for the covering number with respect to $\overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{0.5,p_i}}$. If, instead of setting $B = 1$, we wanted to consider $B = 0.5$ as a bound on the domain, the covering number bound would become only tighter in terms of constant factors. Considering the above facts and applying Theorem 23 recursively, we can write that

$$\begin{aligned}
&\log N_U\left(k\epsilon, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{F}}, \infty, d_{TV}^{\infty}, \overline{\mathcal{X}_{0.5,p_0}}\right) \\
&\le \sum_{i=0}^{k-1} \log N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_i, p_{i+1}], \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{X}_{0.5,p_i}}\right) \\
&\le \sum_{i=0}^{k-1} p_i(p_{i+1}+1) \log\left(30 \frac{p_i^{5/2}\sqrt{\ln\left((5p_i - \epsilon\sigma)/(\epsilon\sigma)\right)}}{\epsilon^{3/2}\sigma^2} \ln\left(\frac{5p_i}{\epsilon\sigma}\right)\right).
\end{aligned}$$

We can now set $\epsilon' = \epsilon/k$ and rewrite the above equation as

$$\begin{aligned}
&\log N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{F}}, \infty, d_{TV}^{\infty}, \overline{\mathcal{X}_{0.5,p_0}}\right) \\
&\le \sum_{i=0}^{k-1} p_i(p_{i+1}+1) \max_i \left\{\log\left(30 \frac{p_i^{5/2}\sqrt{\ln\left((5p_i - \epsilon'\sigma)/(\epsilon'\sigma)\right)}}{\epsilon'^{3/2}\sigma^2} \ln\left(\frac{5p_i}{\epsilon'\sigma}\right)\right)\right\} \\
&\le w \max_i \left\{\log\left(30 \frac{p_i^{5/2}\sqrt{\ln\left((5p_i - \epsilon'\sigma)/(\epsilon'\sigma)\right)}}{\epsilon'^{3/2}\sigma^2} \ln\left(\frac{5p_i}{\epsilon'\sigma}\right)\right)\right\} \\
&\le w \max_i \left\{\log\left(30 \frac{p_i^{5/2}\sqrt{\ln\left(5p_i/(\epsilon'\sigma)\right)}}{\epsilon'^{3/2}\sigma^2} \ln\left(\frac{5p_i}{\epsilon'\sigma}\right)\right)\right\} \\
&\le w \max_i \left\{\log\left(30 \frac{p_i^{5/2}\sqrt{5p_i/(\epsilon'\sigma)}}{\epsilon'^{3/2}\sigma^2} \ln\left(\frac{5p_i}{\epsilon'\sigma}\right)\right)\right\} \\
&\le w \max_i \left\{\log\left(30\sqrt{5} \frac{p_i^3}{\epsilon'^2\sigma^{3/2}} \ln\left(\frac{5p_i}{\epsilon'\sigma}\right)\right)\right\}.
\end{aligned}$$

Using the fact that $\epsilon, \sigma < 1$, we can simplify the above equation and write that

$$\begin{aligned}
&\log N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{F}}, \infty, d_{TV}^{\infty}, \overline{\mathcal{X}_{0.5,p_0}}\right) \\
&\le w \max_i \left\{\log\left((30\sqrt{5})^3 \frac{p_i^3}{\epsilon'^3\sigma^3} \left(\ln\left(\frac{5p_i}{\epsilon'\sigma}\right)\right)^3\right)\right\} \\
&\le w \max_i \left\{3\log\left(30\sqrt{5} \frac{p_i}{\epsilon'\sigma} \ln\left(\frac{5p_i}{\epsilon'\sigma}\right)\right)\right\} \\
&\le w \max_i \left\{3\log\left(30\sqrt{5} \frac{kp_i}{\epsilon\sigma} \ln\left(\frac{5kp_i}{\epsilon\sigma}\right)\right)\right\} \\
&\le w \left(3\log\left(30\sqrt{5} \frac{w^2}{\epsilon\sigma} \ln\left(\frac{5w^2}{\epsilon\sigma}\right)\right)\right) \\
&= O\left(w \log\left(\frac{w}{\epsilon\sigma} \ln\left(\frac{w}{\epsilon\sigma}\right)\right)\right) = \widetilde{O}\left(w \log\left(\frac{1}{\epsilon\sigma}\right)\right),
\end{aligned}$$

where we used the fact that $k \le w$ and $p_i \le w$ for every $0 \le i \le k$. Now that we found an upper bound on the covering number of $\overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{F}}$ for a choice of $p_1, \dots, p_{k-1}$, we can bound the covering number of $\overline{\mathcal{G}_\sigma} \circ \overline{\text{MNET}_\sigma}[p_0, p_k, w]$. The number of different choices that we can have for $p_1, \dots, p_{k-1}$ is at most w^{k-1} since we know that $\sum_{i=1}^{k} p_i p_{i-1} = w$ and therefore $p_i < w$ for every $0 \le i \le k$. Therefore, we can simply take a union of the covering sets for each choice of $p_0, \dots, p_{k-1}$ as a covering set for $\overline{\mathcal{G}_\sigma} \circ \overline{\text{MNET}_\sigma}[p_0, p_k, w]$, which yields to the following covering number bound.

$$\begin{aligned}
&\log N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ \overline{\text{MNET}_\sigma}[p_0, p_k, w], \infty, d_{TV}^{\infty}, \overline{\mathcal{X}_{0.5,p_0}}\right) \\
&\le \log w^k . N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{F}}, \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_{\sigma,p_0}} \circ \overline{\mathcal{X}_{0.5,p_0}}\right) \\
&\le w \log w + \log N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{F}}, \infty, d_{TV}^{\infty}, \overline{\mathcal{G}_{\sigma,p_0}} \circ \overline{\mathcal{X}_{0.5,p_0}}\right) && (k \le w) \\
&= O\left(w \log w + w \log\left(\frac{w}{\epsilon\sigma} \ln\left(\frac{w}{\epsilon\sigma}\right)\right)\right) = O\left(w \log\left(\frac{w}{\epsilon\sigma} \ln\left(\frac{w}{\epsilon\sigma}\right)\right)\right) = \widetilde{O}\left(w \log\left(\frac{1}{\epsilon\sigma}\right)\right),
\end{aligned}$$

□

D.3 Proof of Theorem 26

Proof. We know that

$$\overline{\text{MNET}_\sigma}[p_0, p_k, w] = \bigcup \text{NET}[p_{k-1}, p_k] \circ \ldots \circ \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_1, p_2] \circ \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_0, p_1] \circ \overline{\mathcal{G}_\sigma}.$$

Define $\overline{\mathcal{F}} = \bigcup \text{NET}[p_{k-1}, p_k] \circ \ldots \circ \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_1, p_2] \circ \overline{\mathcal{G}_\sigma} \circ \text{NET}[p_0, p_1]$ and note that $\overline{\mathcal{F}} \circ \overline{\mathcal{G}_\sigma} = \overline{\mathcal{F}_\sigma} = \overline{\text{MNET}_\sigma}[p_0, p_k, w]$. Therefore, we can use Theorem 24 to write that

$$\begin{aligned}
&N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ \text{REC}[\overline{\text{MNET}_\sigma}[p_0, p_k, w], T], \infty, d_{TV}^{\infty}, \overline{\Delta_{0.5, p\times T}}\right) \\
&= N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ \text{REC}[\overline{\mathcal{F}_\sigma}, T], \infty, d_{TV}^{\infty}, \overline{\Delta_{0.5, p\times T}}\right) \\
&\leq N_U\left(\frac{\epsilon}{T}, \overline{\mathcal{G}_\sigma} \circ \overline{\mathcal{F}_\sigma}, \infty, d_{TV}^{\infty}, \overline{\mathcal{X}_{0.5, s}}\right) \\
&= N_U\left(\frac{\epsilon}{T}, \overline{\mathcal{G}} \circ \overline{\text{MNET}_\sigma}[p_0, p_k, w], \infty, d_{TV}^{\infty}, \overline{\mathcal{X}_{0.5, s}}\right).
\end{aligned}$$

We know of a bound on the covering number of $\overline{\mathcal{G}_\sigma} \circ \overline{\text{MNET}_\sigma}[p_0, p_k, w]$ from Theorem 38. Using this bound we can rewrite the above equation as

$$\begin{aligned}
&N_U\left(\epsilon, \overline{\mathcal{G}_\sigma} \circ \text{REC}[\overline{\text{MNET}_\sigma}[p_0, p_k, w], T], \infty, d_{TV}^{\infty}, \overline{\Delta_{0.5, p\times T}}\right) \\
&\leq N_U\left(\frac{\epsilon}{T}, \overline{\mathcal{G}} \circ \overline{\text{MNET}_\sigma}[p_0, p_k, w], \infty, d_{TV}^{\infty}, \overline{\mathcal{X}_{0.5, s}}\right) \\
&= O\left(w \log\left(\frac{wT}{\epsilon\sigma} \ln\left(\frac{wT}{\epsilon\sigma}\right)\right)\right) = \widetilde{O}\left(w \log\left(\frac{T}{\epsilon\sigma}\right)\right).
\end{aligned}$$

□

D.4 Proof of Theorem 19

Proof. From Theorem 43 we can write that

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(\hat{f}, x, y)\right] \\
&\leq \inf_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(f, x, y)\right] + 2 \inf_{\epsilon\in[0,1/2]} \left\{2\left[4\epsilon + \frac{12}{\sqrt{m}} \int_{\epsilon}^{1/2} \sqrt{\ln N_U(\gamma\nu, \mathcal{F}, m, \|.\|_2^{\ell_2})}\, d\nu\right]\right\} + 6\sqrt{\frac{\ln(2/\delta)}{2m}} \\
&\leq \inf_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(f, x, y)\right] + 2\left[8\epsilon + \frac{24}{\sqrt{m}} \int_{\epsilon}^{1/2} \sqrt{\ln N_U(\gamma\nu, \mathcal{F}, m, \|.\|_2^{\ell_2})}\, d\nu\right] + 6\sqrt{\frac{\ln(2/\delta)}{2m}} \qquad (\forall\epsilon\in[0,1/2]) \\
&\leq \inf_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(f, x, y)\right] + 16\epsilon + \frac{24}{\sqrt{m}} \sqrt{\ln N_U(\gamma\epsilon, \mathcal{F}, m, \|.\|_2^{\ell_2})} + 6\sqrt{\frac{\ln(2/\delta)}{2m}},
\end{aligned}$$

where we have used the fact that the integral is over $[0, 1/2]$ and the covering number decreases monotonically with ϵ. □

D.5 Proof of Theorem 15

We are now ready to state the proof of the upper bound on the sample complexity of PAC learning noisy recurrent neural networks with respect to the ramp loss.

Proof. From Theorem 19 we know that if we choose algorithm $\mathcal{A}$ such that for every distribution over $[-1/2, 1/2]^{p\times T} \times \{-1, 1\}$ and any input S of m i.i.d. samples from $\mathcal{D}$ it outputs $\mathcal{A}(S) = \hat{h} = \arg\min_{h\in\mathcal{H}_w} \frac{1}{|S|} \sum_{(x,y)\in S} l_\gamma(h, x, y)$, then with probability at least $1-\delta$ we have

$$\begin{aligned}
&\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma(\hat{h}, U, y)\right] \\
&\leq \inf_{h\in\mathcal{H}_w} \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma(h, U, y)\right] + 16\epsilon + \frac{24}{\sqrt{m}} \sqrt{\log N_U(\gamma\epsilon, \mathcal{H}_w, m, \|.\|_2^{\ell_2})} + 6\sqrt{\frac{\log(2/\delta)}{2m}}.
\end{aligned} \tag{9}$$

We know that $\overline{\mathcal{Q}_w}$ is a class of functions from $[-1/2, 1/2]^{p\times T}$ to $[-1/2, 1/2]$. We also know that $\|x\|_2^{\ell_2} \leq \|x\|_2^{\infty}$ for every x. We can now use Theorem 22 to turn the bound on the covering number of $\overline{\mathcal{Q}_w}$ to a bound on the covering number of $\mathcal{E}(\overline{\mathcal{Q}_w})$. Note that Theorem 22 is stated for functions with outputs in $[-B, B]$ and $\overline{\mathcal{Q}_w} = \overline{\mathcal{G}_\sigma} \circ \text{REC}[\overline{\text{MNET}_\sigma}[p_0, p_k, w], T]$ outputs values in $\overline{\mathcal{G}_{\sigma,p_k}} \circ \overline{\mathcal{X}_{0.5,p_k}}$. However, $\overline{\mathcal{G}_{\sigma,p_k}}$ is a class of zero mean Gaussian random variables that are independent of the output of $\text{REC}[\overline{\text{MNET}_\sigma}[p_0, p_k, w], T]$ and, therefore, they do not change the expectation and the covering number bound for $\mathcal{E}(\overline{\mathcal{Q}_w})$ would be the same as the covering number bound for $\mathcal{E}\left(\text{REC}[\overline{\text{MNET}_\sigma}[p_0, p_k, w], T]\right)$. Thus we know that

$$N_U(\gamma\epsilon, \mathcal{H}_w, m, \|.\|_2^{\ell_2}) \leq N_U(\gamma\epsilon, \mathcal{H}_w, m, \|.\|_2^{\infty}) \leq N_U(\gamma\epsilon, \overline{\mathcal{Q}_w}, \infty, d_{TV}^{\infty}, \overline{\Delta_{0.5,p\times T}}).$$

We can, therefore, rewrite Equation 9 as follows.

$$\begin{aligned}&\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma(\hat{h}, U, y)\right]\\ &\leq \inf_{h\in\mathcal{H}_w} \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma(h, U, y)\right] + 16\epsilon + \frac{24}{\sqrt{m}}\sqrt{\log N_U(\gamma\epsilon, \overline{\mathcal{Q}_2}, \infty, d_{TV}^{\infty}, \overline{\Delta_{0.5,p\times T}})} + 6\sqrt{\frac{\log(2/\delta)}{2m}}.\end{aligned} \tag{10}$$

Therefore, if we find m such that $\frac{1}{\sqrt{m}}\sqrt{\log N_U(\gamma\epsilon, \overline{\mathcal{Q}_w}, \infty, d_{TV}^{\infty}, \overline{\Delta_{p\times T}})} = O(\epsilon)$ and $\sqrt{\frac{\log(1/\delta)}{m}} = O(\epsilon)$ then we can guarantee $\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma(\hat{h}, U, y)\right] \leq \inf_{h\in\mathcal{H}_w} \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma(h, U, y)\right] + O(\epsilon)$

We know of a covering number bound for $\overline{\mathcal{Q}_w}$ from Theorem 26 which is as follows.

$$\log N_U\left(\epsilon, \overline{\mathcal{Q}_w}, \infty, d_{TV}^{\infty}, \overline{\Delta_{0.5,p\times T}}\right) = O\left(w \log\left(\frac{wT}{\epsilon\sigma}\ln\left(\frac{wT}{\epsilon\sigma}\right)\right)\right).$$

We can thus write that

$$\sqrt{\frac{\log N_U(\gamma\epsilon, \overline{\mathcal{Q}_w}, \infty, d_{TV}^{\infty}, \overline{\Delta_{0.5,p\times T}})}{m}} = O(\epsilon) \Leftrightarrow m = O\left(\frac{1}{\epsilon^2} w \log\left(\frac{wT}{\epsilon\sigma}\ln\left(\frac{wT}{\epsilon\sigma}\right)\right)\right)$$

Moreover, if we want $\sqrt{\frac{\log(1/\delta)}{m}} = O(\epsilon)$ then we should have $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$. Combining the above results, we can conclude that

$$m_{\mathcal{H}_w}(\epsilon, \delta) = O\left(\frac{w \log\left(\frac{wT}{\epsilon\sigma}\ln\left(\frac{wT}{\epsilon\sigma}\right)\right) + \log(1/\delta)}{\epsilon^2}\right) = \widetilde{O}\left(\frac{w\log\left(\frac{T}{\sigma}\right) + \log(1/\delta)}{\epsilon^2}\right).$$

samples is sufficient to conclude that with probability at least $1-\delta$ we have $\mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma(\hat{h}, U, y)\right] \leq \inf_{h\in\mathcal{H}_w} \mathbb{E}_{(U,y)\sim\mathcal{D}}\left[l_\gamma(h, U, y)\right] + O(\epsilon)$, which implies PAC learning $\mathcal{H}_w$ with respect to ramp loss with a sample complexity of $m_{\mathcal{H}_w}(\epsilon, \delta)$. ☐

E PAC learning and covering number bounds

In this section, we discuss how we can find a bound on the sample complexity of PAC learning a class of functions with respect to ramp loss from a bound on its covering number. Particularly, we show how to use a bound on covering number to find the number of samples required to ensure uniform convergence with respect to ramp loss. We then connect the uniform convergence results to PAC learning and find the minimum number of samples required to guarantee PAC learning with respect to ramp loss.

We start by defining uniform convergence (with respect to ramp loss).

Definition 39 (Uniform Convergence with Respect to Ramp Loss)**.** *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. We say that $\mathcal{F}$ has uniform convergence property with respect to ramp loss with margin parameter $\gamma > 0$ if there exists some function $m : (0,1)^2 \to \mathbb{N}$ such that for every distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, 1\}$ and every $\epsilon, \delta \in (0,1)$, if S is a set of $m(\epsilon, \delta)$ i.i.d. samples from $\mathcal{D}$, then with probability at least $1-\delta$ (over the randomness of S) for every function $f \in \mathcal{F}$ we have $\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(f, x, y)\right] - \frac{1}{|S|}\sum_{(x,y)\in S} l_\gamma(f, x, y)\right| \leq \epsilon$.*

The *sample complexity* of uniform convergence for class $\mathcal{F}$ is denoted by $m^{\text{UC}}_{\mathcal{F}}(\epsilon,\delta)$, which is the minimum number of samples required to guarantee uniform convergence for $\mathcal{F}$. We now show that uniform convergence implies PAC learning (with respect to ramp loss).

Lemma 40. *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$ that satisfies uniform convergence property with respect to ramp loss. Then for any $(\epsilon,\delta)\in(0,1)$, we have $m_{\mathcal{F}}(\epsilon,\delta)\leq m^{UC}_{\mathcal{F}}(\epsilon/2,\delta)$, i.e., there exists an algorithm $\mathcal{A}$ such that for any distribution $\mathcal{D}$ over $\mathcal{X}\times\{-1,1\}$ and any $(\epsilon,\delta)\in(0,1)$, if S is a set of $m\geq m^{UC}_{\mathcal{F}}(\epsilon/2,\delta)$ i.i.d. samples from $\mathcal{D}$, then with probability at least $1-\delta$, we have that $\mathbb{E}\left[(x,y)\sim\mathcal{D}\right] l_\gamma(\mathcal{A}(S),x,y)\leq \inf_{f\in\mathcal{F}}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(f,x,y)\right]+\epsilon$.*

Proof. Let $\mathcal{A}$ be an algorithm that outputs the function in $\mathcal{F}$ that has the minimum empirical loss, i.e., $\mathcal{A}(S)=\arg\min_{f\in\mathcal{F}}\frac{1}{|S|}\sum_{(x,y)\in S} l_\gamma(f,x,y)$. Since S is a set of $m\geq m^{\text{UC}}_{\mathcal{F}}(\epsilon/2,\delta)$ samples, we know that with probability at least $1-\delta$ we have $\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(f,x,y)\right]-\frac{1}{|S|}\sum_{(x,y)\in S} l_\gamma(f,x,y)\right|\leq \epsilon/2$ for every $f\in\mathcal{F}$. Let $\hat{f}=\mathcal{A}(S)$. Then for every $f\in\mathcal{F}$ we can write that

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(\hat{f},x,y)\right]\leq\frac{1}{|S|}\sum_{(x,y)\in S} l_\gamma(\hat{f},x,y)+\frac{\epsilon}{2}\leq\frac{1}{|S|}\sum_{(x,y)\in S} l_\gamma(f,x,y)+\frac{\epsilon}{2}$$
$$\leq\mathbb{E}_{(x,y)\in\mathcal{D}}\left[l_\gamma(f,x,y)\right]+\frac{\epsilon}{2}+\frac{\epsilon}{2}=\mathbb{E}_{(x,y)\in\mathcal{D}}\left[l_\gamma(f,x,y)\right]+\epsilon.$$

This implies that with $m\geq m^{\text{UC}}_{\mathcal{F}}(\epsilon/2,\delta)$ i.i.d. samples we can guarantee PAC learning with respect to ramp loss with parameters ϵ and δ. In other words, we have $m_{\mathcal{F}}(\epsilon,\delta)\leq m^{\text{UC}}_{\mathcal{F}}(\epsilon/2,\delta)$. □

The following theorem tells us that we can relate the bound on the covering number of a class of functions to the uniform convergence property for that class. The proof relies on bounding the Rademacher complexity of the class by a bound on its covering number (Dudley, 2010) and then relating the bound on the Rademacher complexity to uniform convergence property. See Shalev-Shwartz and Ben-David (2014) and Mohri et al. (2018) for a more detailed discussion and proof.

Theorem 41. *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$ and $\mathcal{F}_\gamma=\{f_\gamma:\mathcal{X}\times\{-1,1\}\to[0,1]\mid f_\gamma(x,y)=r_\gamma\left(-f(x).y\right),f\in\mathcal{F}\}$ be the class of its composition with ramp loss. Let $\mathcal{D}$ be a distribution over $\mathcal{X}\times\{-1,1\}$ and $S\sim\mathcal{D}^m$ be an i.i.d. sample of size m. Then, for any $\delta\in(0,1)$ with probability at least $1-\delta$ (over the randomness of S) for every $f\in\mathcal{F}$ we have*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(f,x,y)\right]$$
$$\leq\frac{1}{|S|}\sum_{(x,y)\in S} l_\gamma(f,x,y)+\inf_{\epsilon\in[0,1/2]}\left\{2\left[4\epsilon+\frac{12}{\sqrt{m}}\int_\epsilon^{1/2}\sqrt{\log N_U(\nu,\mathcal{F}_\gamma,m,\|.\|_2^{\ell_2})}\,d\nu\right]\right\}+3\sqrt{\frac{\log(2/\delta)}{2m}}.$$

It is only left to find a bound on the covering number of $\mathcal{F}_\gamma$ from a bound on the covering number of $\mathcal{F}$. The following lemma helps us finding this bound.

Lemma 42 (From Covering Number of $\mathcal{F}$ to Covering Number of $\mathcal{F}_\gamma$)**.** *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$ and $\mathcal{F}_\gamma=\{f_\gamma:\mathcal{X}\times\{-1,1\}\to[0,1]\mid f_\gamma(x,y)=r_\gamma\left(-f(x).y\right),f\in\mathcal{F}\}$ be the class of its composition with ramp loss. Then we have*

$$N_U(\epsilon,\mathcal{F}_\gamma,m,\|.\|_2^{\ell_2})\leq N_U(\gamma\epsilon,\mathcal{F},m,\|.\|_2^{\ell_2}).$$

Proof. First, it is easy to verify that r_γ (with respect to the first input) is a Lipschitz continuous function with respect to $\|.\|_2$ with Lipschitz factors of $1/\gamma$; see e.g., section A.2 in Bartlett et al. (2017).

Fix an input set $S=\{(x_1,y_1),\ldots,(x_m,y_m)\}\subset\mathcal{X}\times\mathcal{Y}$ and let $C=\{\hat{f}_{i|S}\mid\hat{f}_i\in\mathcal{F},\,i\in[r]\}$ be an $(\gamma\epsilon)$-cover for $\mathcal{F}_{|S}$. For the simplicity of notation, we denote the composition of $\hat{f}_i$ with ramp loss by $\hat{f}_{\gamma,i}$. Now, we prove that $C_\gamma=\{\hat{f}_{\gamma,i_{|S}}\mid\hat{f}_{\gamma,i}\in\mathcal{F}_\gamma,\,i\in[r]\}$ is also an ϵ-cover for $\mathcal{F}_{\gamma_{|S}}$.

Given any $f\in\mathcal{F}$, there exists $\hat{f}_{i|S}\in C$ such that

$$\left\|(\hat{f}_i(x_1),\ldots,\hat{f}_i(x_m))-(f(x_1),\ldots,f(x_m))\right\|_2^{\ell_2}\leq\gamma\epsilon.$$

We can then write that

$$
\begin{aligned}
&\left\|(\hat{f}_{\gamma,i}(x_1),\ldots,\hat{f}_{\gamma,i}(x_m))-(f_\gamma(x_1),\ldots,f_\gamma(x_m))\right\|_2^{\ell_2}\\
&=\sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(\hat{f}_{\gamma,i}(x_k)-f_\gamma(x_k)\right)^2}\\
&=\sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(r_\gamma\left(-\hat{f}_i(x_k).y_k\right)-r_\gamma(-f(x_k).y_k)\right)^2}.
\end{aligned}
\tag{11}
$$

From the Lipschitz continuity of $r_\gamma(x)$ we can conclude that for any $(x,y)\in\mathcal{X}\times\mathcal{Y}$,

$$
\left|r_\gamma(-f(x).y)-r_\gamma(-\hat{f}_i(x).y)\right|\leq\frac{1}{\gamma}\left|\hat{f}_i(x)-f(x)\right|.
$$

Taking the above equation into account, we can rewrite Equation 11 as

$$
\begin{aligned}
&\left\|(\hat{f}_{\gamma,i}(x_1),\ldots,\hat{f}_{\gamma,i}(x_m))-(f_\gamma(x_1),\ldots,f_\gamma(x_m))\right\|_2^{\ell_2}\\
&\leq\frac{1}{\gamma}\sqrt{\frac{1}{m}\sum_{k=1}^{m}\left((\hat{f}_i(x_k)-f(x_k))\right)^2}\\
&\leq\frac{1}{\gamma}\left\|(\hat{f}_i(x_1),\ldots,\hat{f}_i(x_m))-(f(x_1),\ldots,f(x_m))\right\|_2^{\ell_2}\\
&\leq\frac{1}{\gamma}\gamma\epsilon\\
&\leq\epsilon.
\end{aligned}
$$

In other words, for any $f_{\gamma_{|S}}\in\mathcal{F}_{\gamma_{|S}}$ there exists $\hat{f}_{\gamma,i_{|S}}\in S$ such that $\left\|\hat{f}_{\gamma,i_{|S}}-f_{\gamma_{|S}}\right\|_2^{\ell_2}\leq\epsilon$ and, therefore, C_γ is an ϵ-cover for $\mathcal{F}_{\gamma_{|S}}$ and the result follows. □

We can now combine Theorem 41, Lemma 40, and Lemma 42 to state the following theorem, which implies that we can relate a bound on the covering number of a class $\mathcal{F}$ to PAC learning $\mathcal{F}$ with respect to ramp loss.

Theorem 43. *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. There exists an algorithm $\mathcal{A}$ with the following property: For every distribution $\mathcal{D}$ over $\mathcal{X}\times\{-1,1\}$ and every $\delta\in(0,1)$, if S is a set of m i.i.d. samples from $\mathcal{D}$, the algorithm outputs a hypothesis $f=\mathcal{A}(S)$ such that with probability at least $1-\delta$ (over the randomness of S and $\mathcal{A}$) we have*

$$
\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(f,x,y)\right]\\
&\leq\inf_{f\in\mathcal{F}}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(f,x,y)\right]+2\inf_{\epsilon\in[0,1/2]}\left\{2\left[4\epsilon+\frac{12}{\sqrt{m}}\int_\epsilon^{1/2}\sqrt{\log N_U(\nu,\mathcal{F}_\gamma,m,\|.\|_2^{\ell_2})}\,d\nu\right]\right\}+6\sqrt{\frac{\log(2/\delta)}{2m}}.
\end{aligned}
$$

Proof. From Theorem 41 we know that for every $f\in\mathcal{F}$ with probability at least $1-\delta$ we have

$$
\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(f,x,y)\right]\\
&\leq\frac{1}{|S|}\sum_{(x,y)\in S}l_\gamma(f,x,y)+\inf_{\epsilon\in[0,1/2]}\left\{2\left[4\epsilon+\frac{12}{\sqrt{m}}\int_\epsilon^{1/2}\sqrt{\log N_U(\nu,\mathcal{F}_\gamma,m,\|.\|_2^{\ell_2})}\,d\nu\right]\right\}+3\sqrt{\frac{\log(2/\delta)}{2m}}.
\end{aligned}
$$

Lemma 40 suggests that if we choose algorithm $\mathcal{A}$ such that $\mathcal{A}(S) = \hat{f} = \arg\min_{f\in\mathcal{F}} \frac{1}{|S|}\sum_{(x,y)\in S} l_\gamma(f,x,y)$ then for any $f \in \mathcal{F}$ with probability at least $1-\delta$ we have

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(\hat{f},x,y)\right]\\
&\leq \frac{1}{|S|}\sum_{(x,y)\in S} l_\gamma(\hat{f},x,y) + \inf_{\epsilon\in[0,1/2]}\left\{2\left[4\epsilon + \frac{12}{\sqrt{m}}\int_{\epsilon}^{1/2}\sqrt{\log N_U(\nu,\mathcal{F}_\gamma,m,\|.\|_2^{\ell_2})}\,d\nu\right]\right\} + 3\sqrt{\frac{\log(2/\delta)}{2m}}\\
&\leq \frac{1}{|S|}\sum_{(x,y)\in S} l_\gamma(f,x,y) + \inf_{\epsilon\in[0,1/2]}\left\{2\left[4\epsilon + \frac{12}{\sqrt{m}}\int_{\epsilon}^{1/2}\sqrt{\log N_U(\gamma\nu,\mathcal{F},m,\|.\|_2^{\ell_2})}\,d\nu\right]\right\} + 3\sqrt{\frac{\log(2/\delta)}{2m}}\\
&\leq \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(f,x,y)\right] + 2\inf_{\epsilon\in[0,1/2]}\left\{2\left[4\epsilon + \frac{12}{\sqrt{m}}\int_{\epsilon}^{1/2}\sqrt{\log N_U(\gamma\nu,\mathcal{F},m,\|.\|_2^{\ell_2})}\,d\nu\right]\right\} + 6\sqrt{\frac{\log(2/\delta)}{2m}}\\
&\leq \inf_{f\in\mathcal{F}}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l_\gamma(f,x,y)\right] + 2\inf_{\epsilon\in[0,1/2]}\left\{2\left[4\epsilon + \frac{12}{\sqrt{m}}\int_{\epsilon}^{1/2}\sqrt{\log N_U(\gamma\nu,\mathcal{F},m,\|.\|_2^{\ell_2})}\,d\nu\right]\right\} + 6\sqrt{\frac{\log(2/\delta)}{2m}}.
\end{aligned}$$

□

In Appendix D we use the above theorem together with an approximation of the right hand side of the above inequality to find an upper bound on the sample complexity of PAC learning noisy recurrent neural networks with respect to ramp loss.

F Missing proof from Section 7

F.1 Proof of Theorem 22

Proof. Let $S = \{U_1,\dots,U_m\} \subset \mathbb{R}^{p\times T}$ be an input set and define $\overline{S} = \{\overline{U_1},\dots,\overline{U_m}\} \subset \overline{\Delta_{p\times T}}$. Let $C = \{\overline{\hat{f}}_{1|\overline{S}},\dots,\overline{\hat{f}}_{r|\overline{S}} \mid \overline{\hat{f}}_r \in \overline{\mathcal{F}}, i\in[r]\}$ be an ϵ-cover for $\overline{\mathcal{F}}_{|\overline{S}}$ with respect to d_{TV}^{∞}. Denote $\mathcal{H} = \mathcal{E}(\overline{\mathcal{F}})$ and let $\hat{\mathcal{H}} = \left\{\hat{h}_i(x) = \mathbb{E}_{\overline{\hat{f}_i}}\left[\,\overline{\hat{f}}_i(x)\right] \mid i\in[r]\right\} \subset \mathcal{E}(\overline{\mathcal{F}})$ be a new set of non-random function.

Given any random function $\overline{f} \in \overline{\mathcal{F}}$ and considering the fact that C is an ϵ-cover for $\overline{\mathcal{F}}_{|\overline{S}}$ we know there exists $\overline{\hat{f}}_i$, $i\in[r]$ such that

$$d_{TV}^{\infty}\left(\overline{\hat{f}}_{i|\overline{S}}, \overline{f}_{|\overline{S}}\right) = d_{TV}^{\infty}\left((\overline{\hat{f}}_i(\overline{U_1}),\dots,\overline{\hat{f}}_i(\overline{U_m})),(\overline{f}(\overline{U_1}),\dots,\overline{f}(\overline{U_m}))\right) \leq \epsilon.$$

From the above equation we can conclude that for any $k\in[m]$ we have $d_{TV}\left(\overline{\hat{f}}_i(\overline{U_k}),\overline{f}(\overline{U_k})\right) \leq \epsilon$. Further, for the corresponding $h,\hat{h}_i \in \mathcal{E}(\overline{\mathcal{F}})$, we know that

$$\begin{aligned}
\hat{h}_i(U_k) &= \mathbb{E}_{\overline{\hat{f}_i}}\left[\,\overline{\hat{f}}_i(\overline{U_k})\right] = \int_{\mathbb{R}^d} x\mathscr{D}(\overline{\hat{f}}_i(\overline{U_k}))(x)dx,\\
h(U_k) &= \mathbb{E}_{\overline{f}}\left[\,\overline{f}(\overline{U_k})\right] = \int_{\mathbb{R}^d} x\mathscr{D}(\overline{f}(\overline{U_k}))(x)dx.
\end{aligned}$$

Denote $I = \mathscr{D}(\overline{f}(\overline{U_k}))$ and $\hat{I} = \mathscr{D}(\overline{\hat{f}}_i(\overline{U_k}))$. Define two new density functions I_{diff} and $\hat{I}_{diff}$ as

$$I_{diff}(x) = \begin{cases} \dfrac{I(x)-\hat{I}(x)}{d_{TV}(I,\hat{I})} & I(x)\geq\hat{I}(x)\\[2ex] 0 & \text{otherwise,}\end{cases}$$

$$\hat{I}_{diff}(x) = \begin{cases} \dfrac{\hat{I}(x)-I(x)}{d_{TV}(I,\hat{I})} & \hat{I}(x)\geq I(x)\\[2ex] 0 & \text{otherwise.}\end{cases}$$

Also, we define I_{min} as

$$I_{min}(x) = \frac{\min\{I(x), \hat{I}(x)\}}{\int \min\{I(x), \hat{I}(x)\}dx} = \frac{\min\{I(x), \hat{I}(x)\}}{1 - d_{TV}(I, \hat{I})}.$$

We can verify that

$$I(x) = \left(1 - d_{TV}(I, \hat{I})\right) I_{min}(x) + d_{TV}(I, \hat{I}).I_{diff}(x)$$
$$\hat{I}(x) = \left(1 - d_{TV}(I, \hat{I})\right) I_{min}(x) + d_{TV}(I, \hat{I}).\hat{I}_{diff}(x).$$

We then find the ℓ_2 distance between $\hat{h}_i(U_k)$ and $h(U_k)$ by

$$\begin{aligned}
&\left\|\hat{h}_i(U_k) - h(U_k)\right\|_2 \\
&= \left\|\int_{\mathbb{R}^d} x\hat{I}(x)dx - \int_{\mathbb{R}^d} xI(x)dx\right\|_2 \\
&= \left\|\int_{\mathbb{R}^d} x\left[\left(1 - d_{TV}(I, \hat{I})\right) I_{min}(x) + d_{TV}(I, \hat{I}).\hat{I}_{diff}(x)\right]\right. \\
&\left.-x\left[\left(1 - d_{TV}(I, \hat{I})\right) I_{min}(x) + d_{TV}(I, \hat{I}).I_{diff}(x)\right] dx\right\|_2 \\
&= \left\|\int_{\mathbb{R}^d} x d_{TV}(I, \hat{I})\left[\hat{I}_{diff}(x) - I_{diff}(x)\right] dx\right\|_2 \\
&= d_{TV}(I, \hat{I})\left\|\int_{\mathbb{R}^d} x\left[\hat{I}_{diff}(x) - I_{diff}(x)\right] dx\right\|_2 \\
&\leq 2B\sqrt{q}\, d_{TV}\left(\overline{f}(\overline{U_k}), \overline{\hat{f}}_i(\overline{U_k})\right) && \text{(Bounded domain } [-B, B]^q \text{ and triangle inequality)} \\
&\leq 2B\epsilon\sqrt{q}.
\end{aligned}$$

Since this result holds for any $k \in [m]$, we have

$$\begin{aligned}
\|\hat{h_{i|S}} - h_{|S}\|_2^{\ell_2} &= \sqrt{\frac{1}{m}\sum_{k=1}^{m}\left\|\hat{h_i}(U_k) - h(U_k)\right\|_2^2} \\
&\leq \sqrt{\frac{1}{m}\sum_{k=1}^{m}(2B\sqrt{q})^2\left(d_{TV}\left(\overline{f}(\overline{U_k}), \overline{\hat{f}}_i(\overline{U_k})\right)\right)^2} \leq 2B\sqrt{q}\sqrt{\frac{1}{m}\sum_{k=1}^{m}\epsilon^2} \leq 2B\epsilon\sqrt{q}.
\end{aligned}$$

Therefore, $\hat{\mathcal{H}}_{|S}$ is a $2B\epsilon\sqrt{q}$ cover for $\mathcal{H}_{|S}$ with respect to $\|.\|_2^{\ell_2}$ and $|\hat{\mathcal{H}}_{|S}| = r$. This holds for any subset S of $\mathbb{R}^{p\times T}$ with $|S| = m$. Therefore,

$$N_U(2B\epsilon\sqrt{q}, \mathcal{E}(\overline{\mathcal{F}}), m, \|.\|_2^{\ell_2}) \leq N_U(\epsilon, \overline{\mathcal{F}}, m, d_{TV}^{\infty}, \overline{\Delta_{p\times T}}) \leq N_U(\epsilon, \overline{\mathcal{F}}, \infty, d_{TV}^{\infty}, \overline{\Delta_{p\times T}}).$$

□